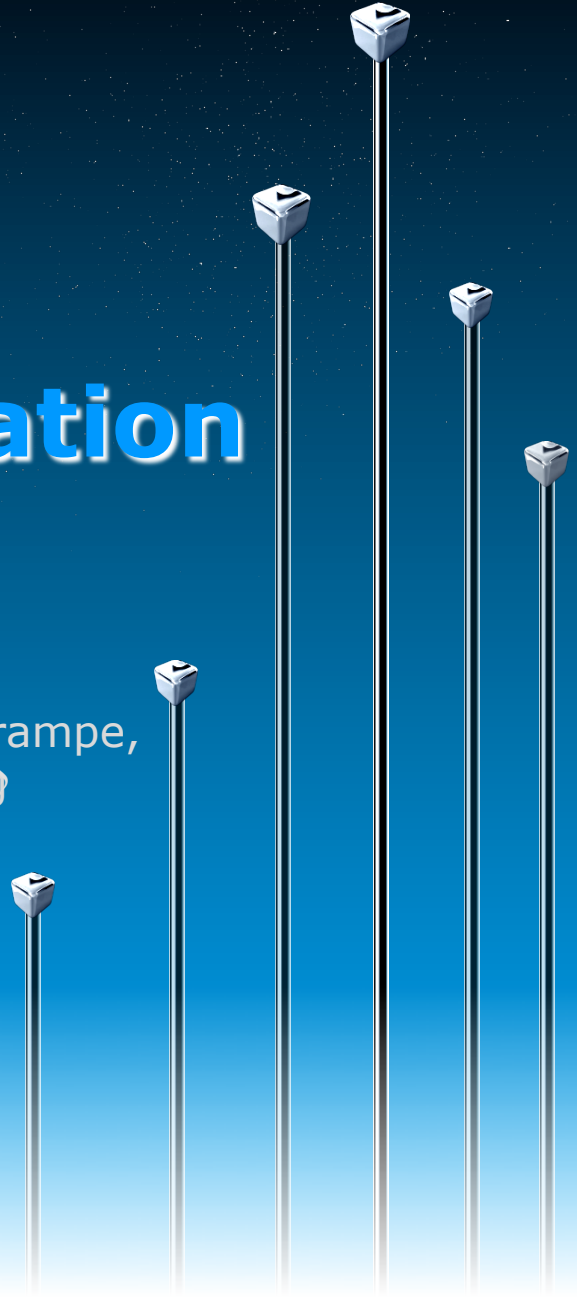


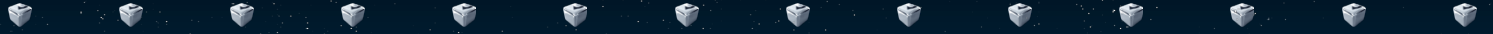
# Demographic Information of TV shows

Team members: Jing Liu, Yiwen Sun, Brandon Trampe,  
Xiaoqian Zhao, Grace Bishop, Yicheng Kang





# Outline



**1**

**Introduction**

**2**

**Simulation with External Data**

**3**

**Ordinary Least Square**

**4**

**Conclusion**



# Introduction

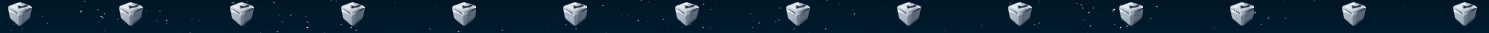


❏ **What does the data look like:**

Description	SubTotal	Clothing	Electronics	HomeFurnish	Housewares	SportingGoods	Toys&Games
ER	19640	31.65	4.22	10.65	7.22	4.97	9.58
Friends	16000	23.92	3.52	7.54	5.86	4.5	5.64
Frasier	14840	31.03	3.98	10.51	8.25	4.66	8.51
Jesse	13550	19.14	3.77	6.49	4.97	4.24	5.28
Ally_McBeal	10190	21.8	2.4	6.42	6.32	4.58	6.67



# Introduction



## **What clients want?**

**Overall objective: demographic information of the audience of the 5 shows.**

**Age**

**Education**

**Gender**

**Employment**

**Marital status**

**Income**



# Introduction

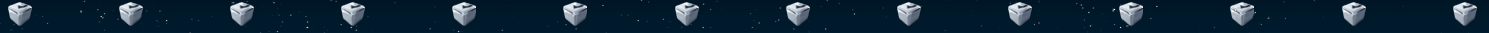


 **More data...** 

Description	SubTotal	Cloth	Elec	Home	House	Sport	Toy
18--24	23965	15.68	2.34	5.05	4.21	4.19	5.27
25--34	42832	22.57	3.98	7.99	5.74	4.2	8.99
35--44	39908	31.02	4.94	11.63	6.85	6.12	10.97
45--54	27327	31.1	5.72	9.43	8.39	4.99	6.54
55--64	21238	28.26	3.22	9.83	8.39	4.17	7.4
65--over	30552	24.36	3.09	7.27	4.66	1.43	4.07
Clg_Grad	36463	36.34	5.97	10.16	6.39	6.24	10.41
Attd_Clg	44294	29.08	4.55	10.92	6.86	4.37	8.28
High_Sch	66741	23.7	3.27	8.1	7.02	4.2	7.32
No_High_Sch	38324	15.14	2.75	5.77	4.3	2.38	4.56
...							



# Introduction



## **Strange...**

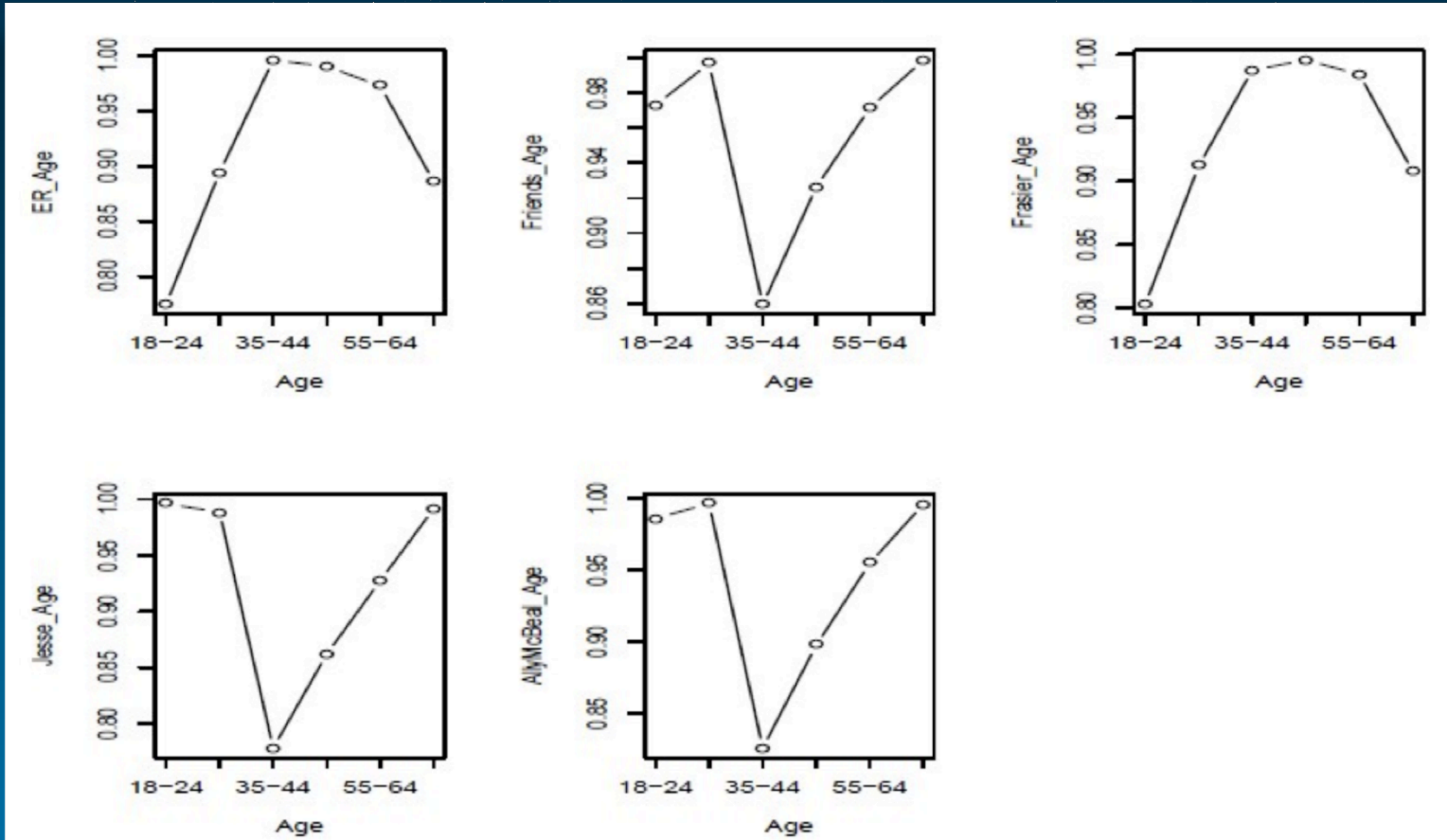
1. The demographic information in the data set is NOT about the audience of the shows.
2. It is collected from a different agency.
3. We need to connect the two sets of information.♪



# Introduction



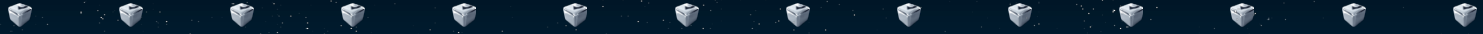
## Data exploratory



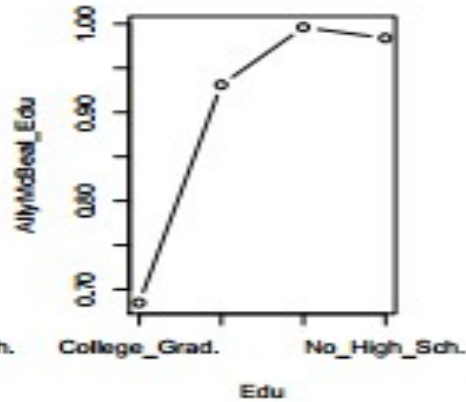
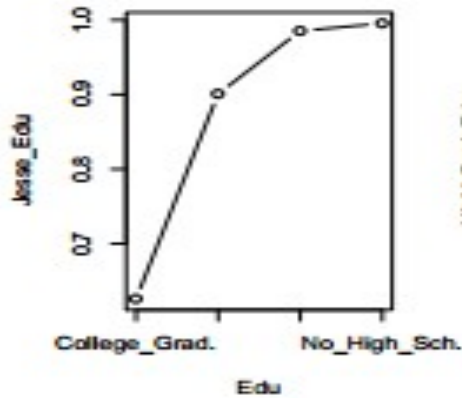
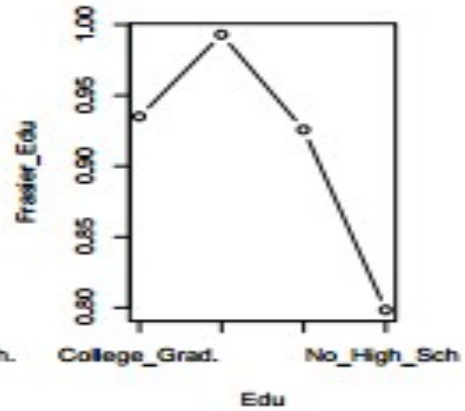
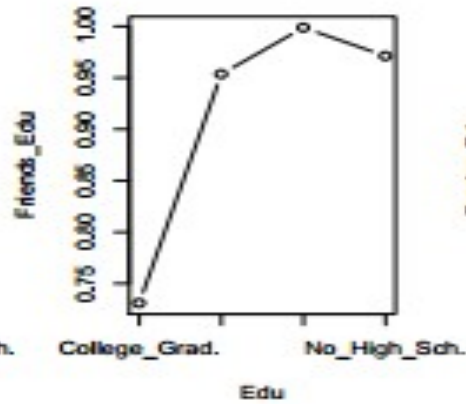
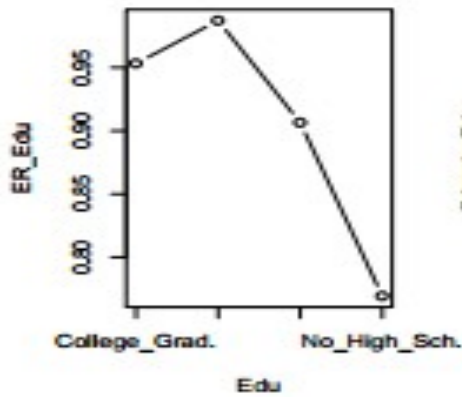
Similar plots can be made for Education, Marital Status, Employment, Income.



# Introduction



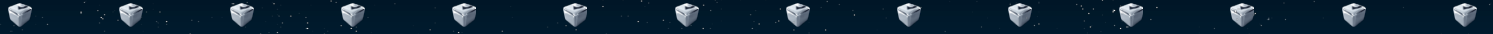
## ▣ Data exploratory



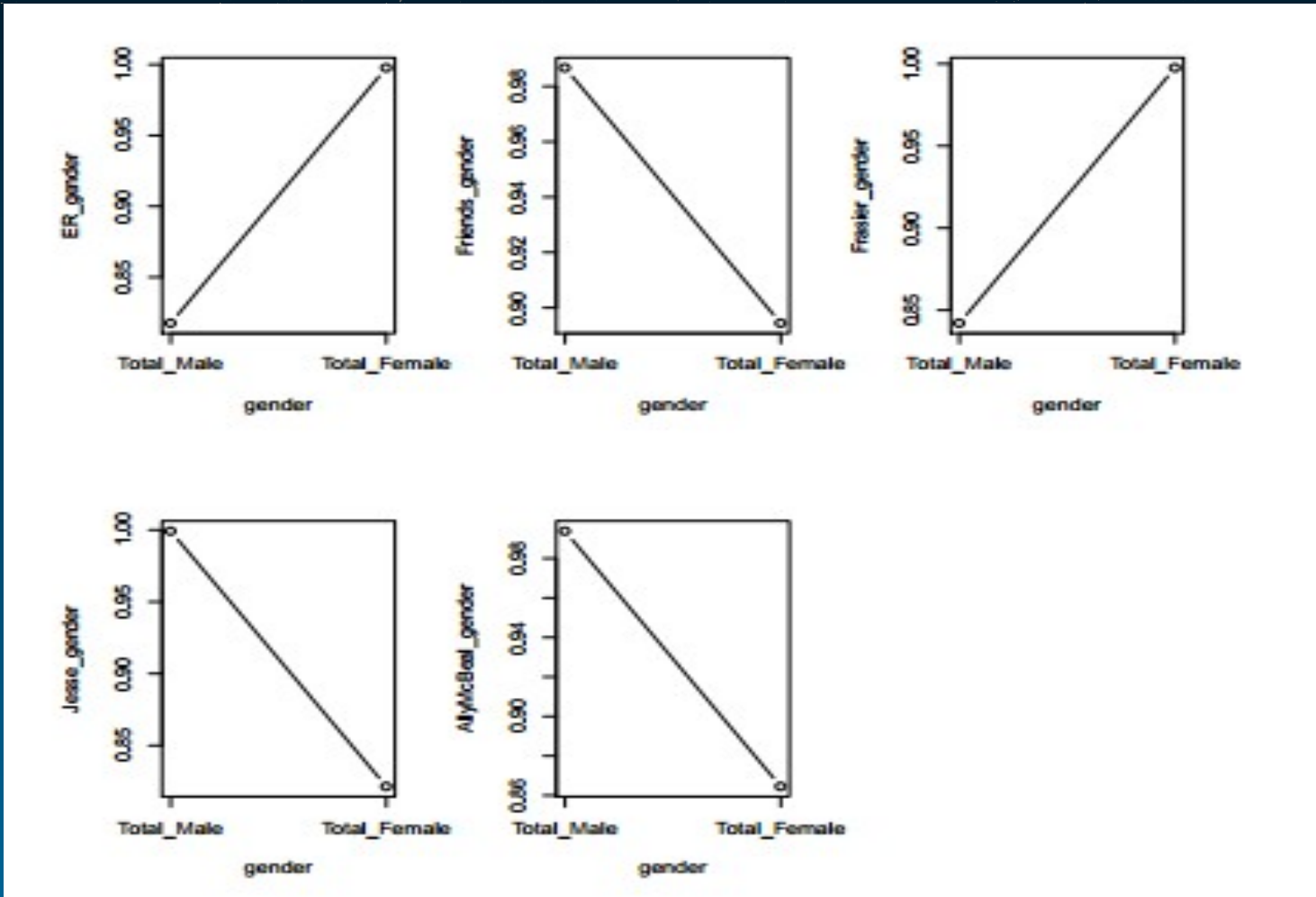




# Introduction



## ▣ Data exploratory ↷

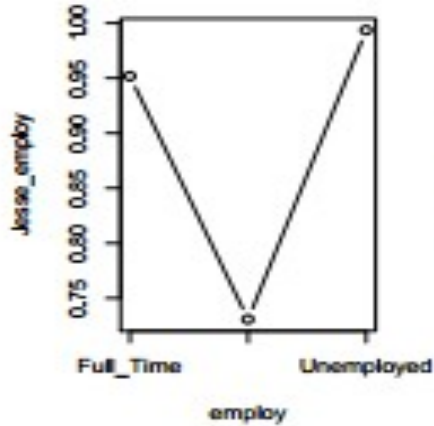
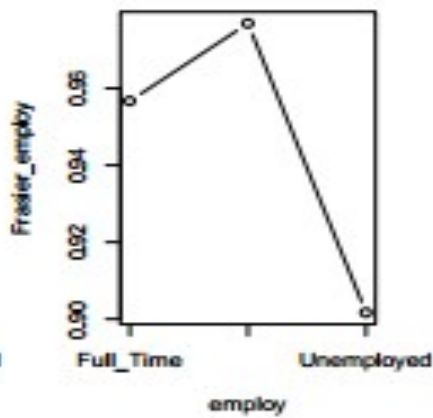
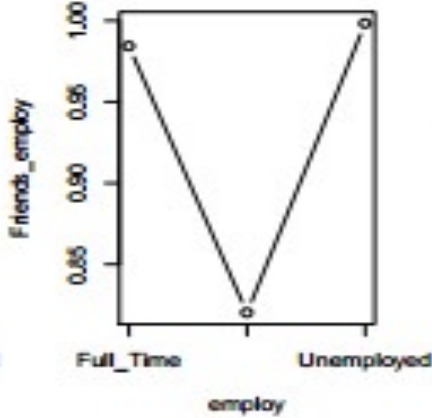




# Introduction



## ▣ Data exploratory

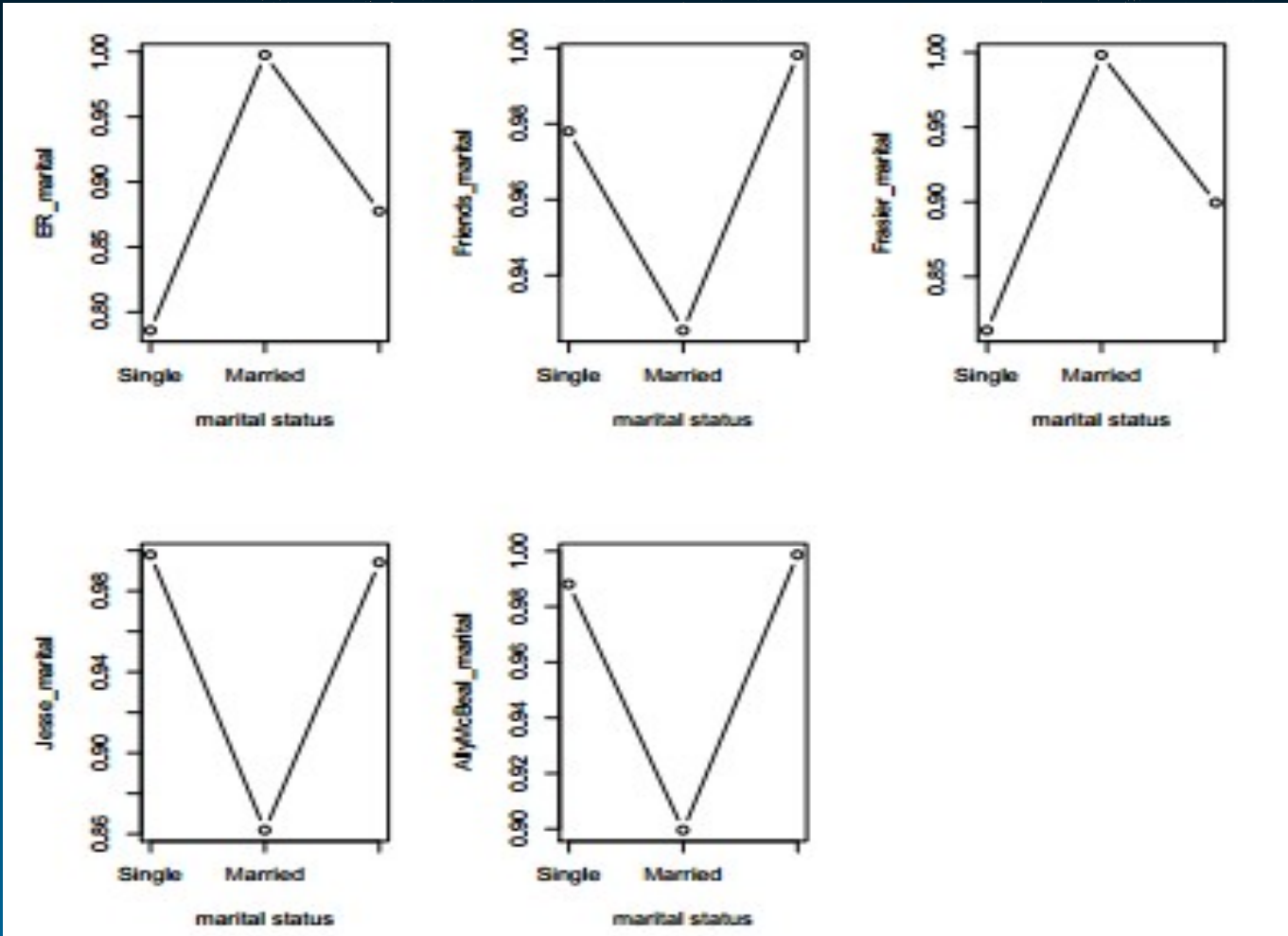




# Introduction



## ▣ Data exploratory ↵

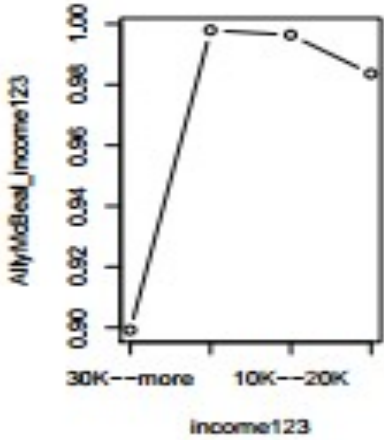
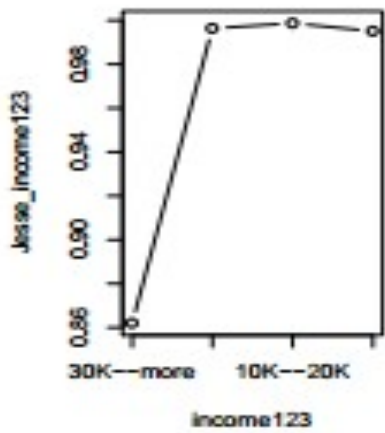
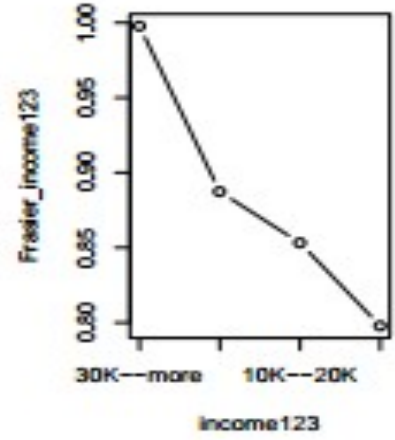
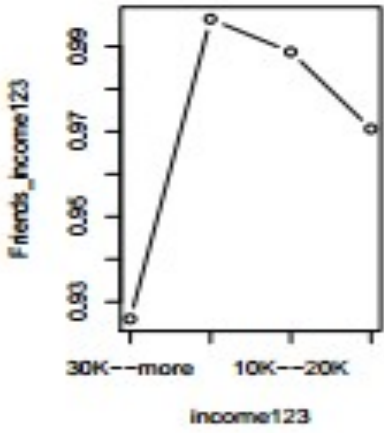
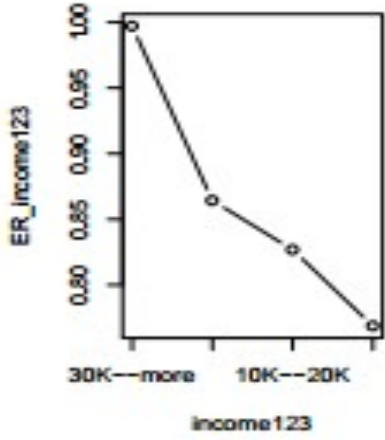




# Introduction

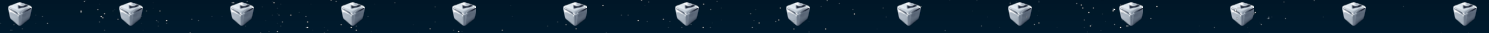


## ▣ Data exploratory





# Simulation with External Data

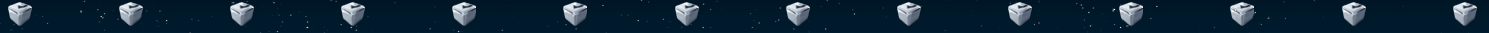


**2**

**Simulation with External Data**



# Simulation with External Data



## Problem

- 1. we only have marginal distributions for each category, but we really want to know joint distributions.**
- 2. We don't have a good way to construct confident intervals.**

**Solution: Create simulated dataset with external data set.**



# Simulation with External Data



## New data set

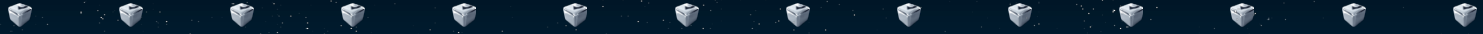
**Aim:** The created data set should support all of the marginal distributions.

**Source:** American Community Survey provides observed joint distributions for all demographic categories.

**Assumption:** Correlations in provided data set are similar to those in all of the United States.›



# Simulation with External Data



## Getting a joint distribution



Create data set

### Requirements to create data:

- the marginal distributions are that data.

### Issue: create one record for each combination in the data and then calculate

- a weight for each case.
- This takes lots of computing power.





# Simulation with External Data



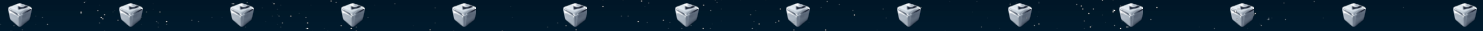
## Getting a joint distribution



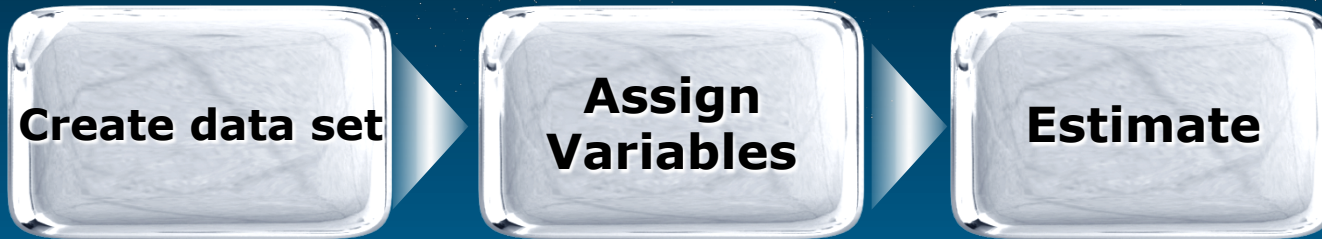
- ❏ **Create dummy variables for purchasing in each category and watching a TV show.**  
**Issue:** This again takes lots of computing power to get all of the marginal distributions correct.
- ❏ **Or, assign a probability of watching a TV show or buying a product for each weighted case.**



# Simulation with External Data



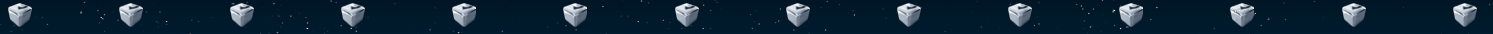
## Getting a joint distribution



- ❖ Create initial 'guess' for each weight based on ACS. Apply to Nielsen data.
- ❖ Adjust weight for Category X such that marginal distribution matches, then for Category Y, Category Z, etc.
- ❖ Each adjustment causes weights in all other categories to be off, so they must be adjusted. Multiple iterations results in 'correct' weights.
- ❖ Then, perform similar, iterative process for watching shows and purchasing products.



# Simulation with External Data



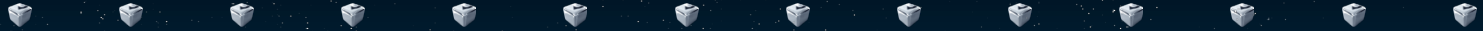
## Getting a joint distribution



- Repeat this process several times to generate a point estimate and standard errors



# Simulation with External Data



**However**♪

**1. The trial for only a couple of demographic categories, for 1 product, and for 1 TV show takes a long time (20 minutes).**

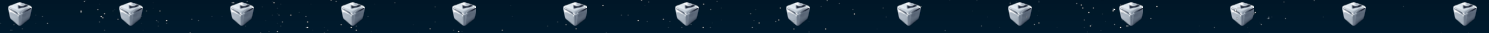
**Ridiculous amount of computing time/power would be needed for the whole data set.**

**2. The estimated demographic distributions that watched each show were almost exactly the same each run.**

**Example: 18,137.28 versus 18,137.31**



# Ordinary Least Square



**3**

**Ordinary Least Square**



# Ordinary Least Square

## Question to answer

**1. To find the conditional probability within each demographic subgroup conditioned on the people who watch one of the five TV shows.**

**(Ex: the probability of ER's audience's salary between 60k to 75k)**

**2. Reversely, if we know a single demographic feature of a person, what is the probability for it to watch each TV program?**

**(Ex: the probability of employed female to watch ER.)**





# Ordinary Least Square



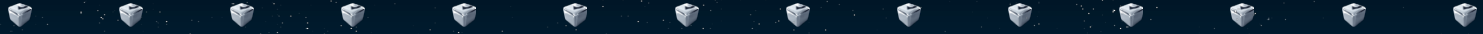
## Process for the 1<sup>st</sup> problem



-  We are dividing the 29 demographic variables into several subgroups, such as age, education, income, etc.
-  From the dataframe, we know how the purchasing behavior is distributed within each of the subgroups and the purchase distribution of each of the five TV shows.



# Ordinary Least Square



Subgroup Description	6 categories
18--24	...
25--34	...
35--44	...
45--54	...
55--64	...
65--over	...
College_Grad.	...
Attend_College	...
High_School	...
No_High_Sch.	...
Total_Male	...
Total_Female	...
Employed_Male	...
Employed_Fem.	...

Full_Time	...
Part_Time	...
Unemployed	...
Single	...
Married	...
Div/Sep/Wid	...
Parents	...
75K--more	...
60K--more	...
50K--more	...
40K--more	...
30K--more	...
20K--29K	...
10K--20K	...
under_10K	...

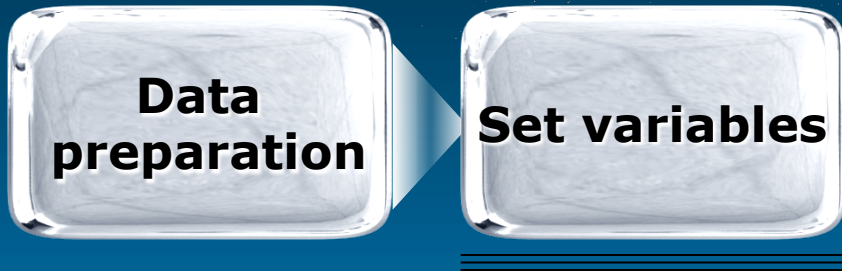




# Ordinary Least Square



## ❖ Process for the 1<sup>st</sup> problem



- ❖ **Without loss of generality, we will demonstrate the estimation between the second employment subgroup and the TV show "ER".**
- ❖ **The second employment subgroup has three categories: full time, part time, and unemployed.**



# Ordinary Least Square



## ❖ Process for the 1<sup>st</sup> problem

	clothing	electronics	home	house	sporting	toys
D1	26.65	4.46	9.2	6.31	8.00	5.03
D2	33.51	2.98	10.74	7.78	5.47	11.51
D3	22.73	3.37	7.41	6.01	2.76	6.2
ER	31.65	4.22	10.65	7.22	4.97	9.58

N1	110363
N2	11047
N3	64412

- ❖ Denote the marginal distribution of the three categories (full time, part time, and unemployed) are D1, D2, D3, respectively.
- ❖ Each of the three categories have  $W_i$  people.



# Ordinary Least Square



## ❖ Process for the 1<sup>st</sup> problem



- ❖  $F = |W_1 * P_1 * D_1 + W_2 * P_2 * D_2 + W_3 * P_3 * D_3 - (W_1 + W_2 + W_3) * TV_1|^2$   
Minimize F where  $P_1 + P_2 + P_3 = 1$ .
- ❖ The objective function is a homogeneous quadratic function of  $P_1, P_2, P_3$ .
- ❖ We have two constraints:  
 $P_1, P_2, P_3$  are all non-negative.  
 $P_1 + P_2 + P_3 = 1$



# Ordinary Least Square



## ❖ Process for the 1<sup>st</sup> problem

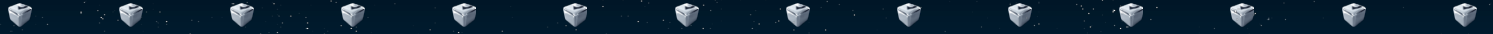
	clothing	electronics	home	house	sporting	toys
D1	26.65	4.46	9.2	6.31	8.00	5.03
D2	33.51	2.98	10.74	7.78	5.47	11.51
D3	22.73	3.37	7.41	6.01	2.76	6.2
ER	31.65	4.22	10.65	7.22	4.97	9.58

N1	110363
N2	11047
N3	64412

W1	0.594
W2	0.059
W3	0.347



# Ordinary Least Square



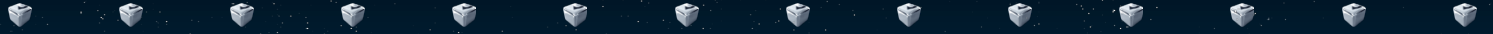
## ❖ Process for the 1<sup>st</sup> problem



- ❖ We could solve this problem by using the "optim" function in R.
- ❖ By minimizing this kind of quadratic target function, we can calculate the estimate of the conditional distribution between every subgroup and every TV show.



# Ordinary Least Square



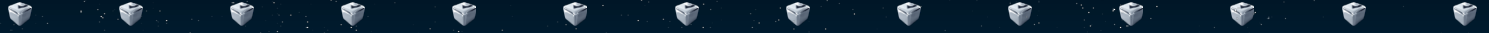
## Process for the 2<sup>nd</sup> problem

The method for the 2<sup>nd</sup> problem is the same as in the 1<sup>st</sup> problem.

The only difference is we assign  $P_i$  to 5 TV shows under each demographic category.



# Conclusion

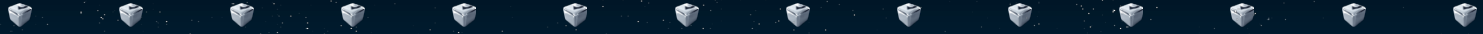


4

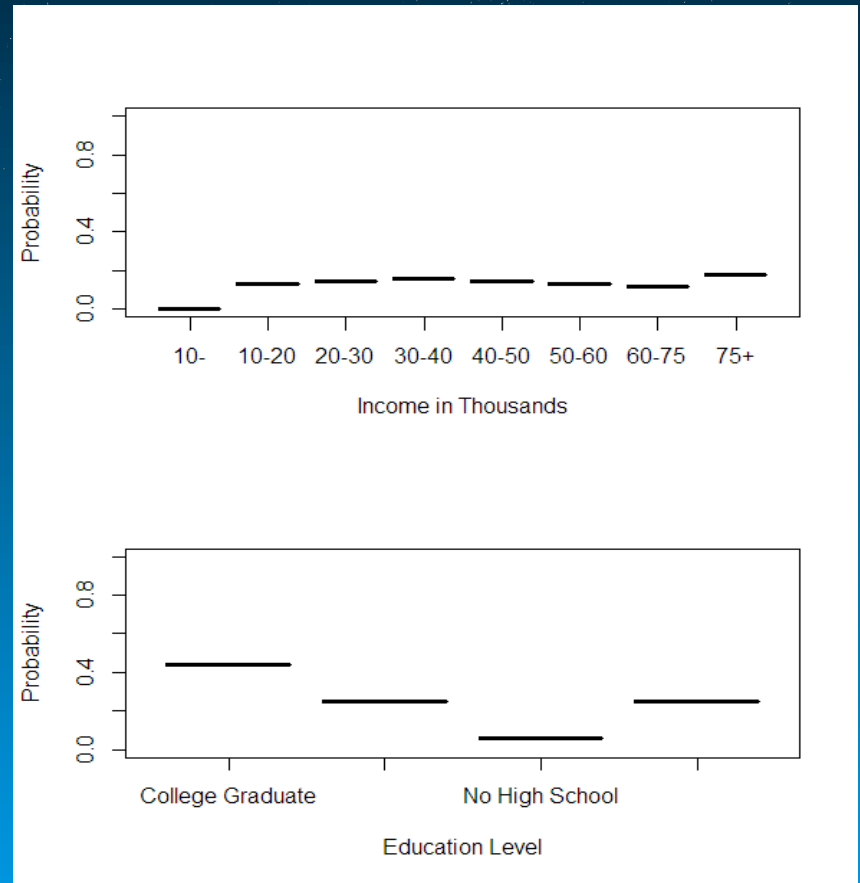
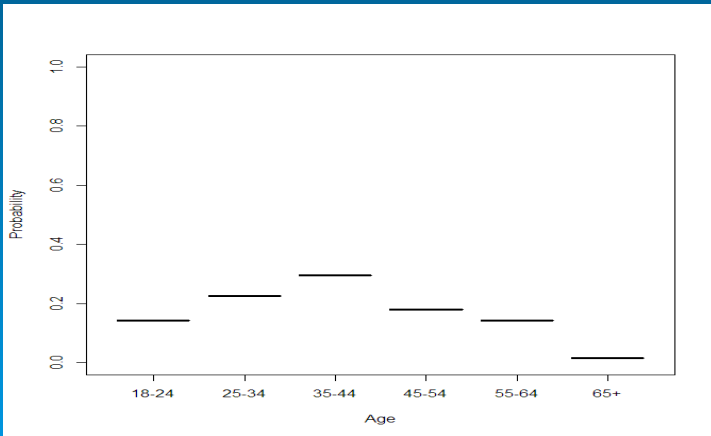
**Conclusion**



# Conclusion



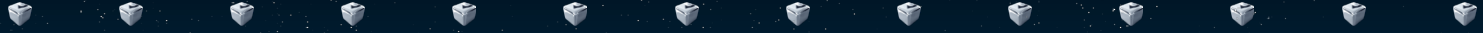
## E.R.



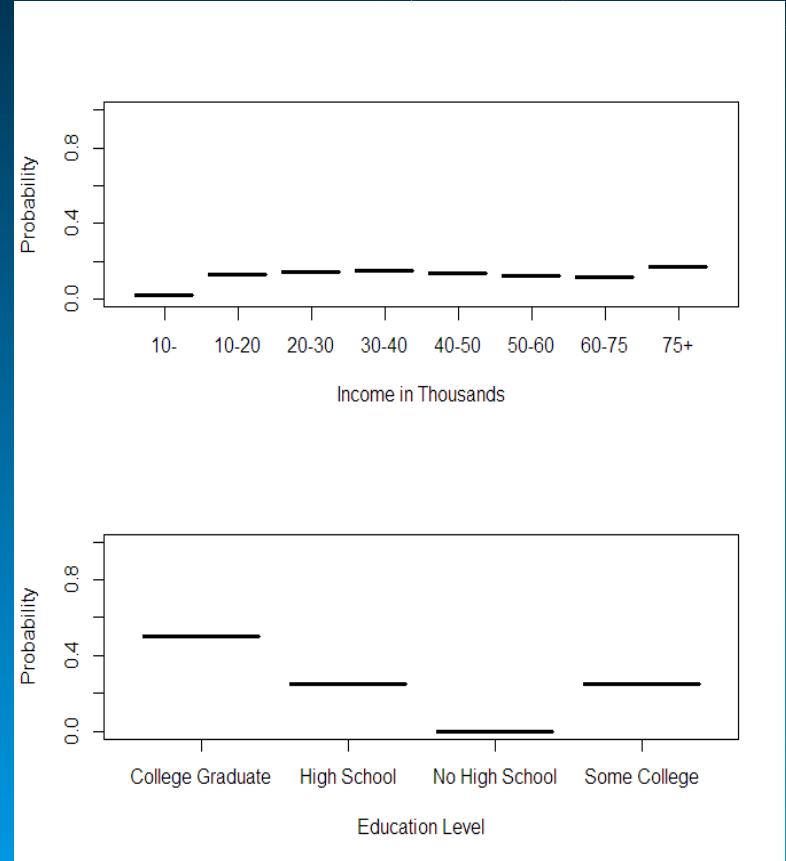
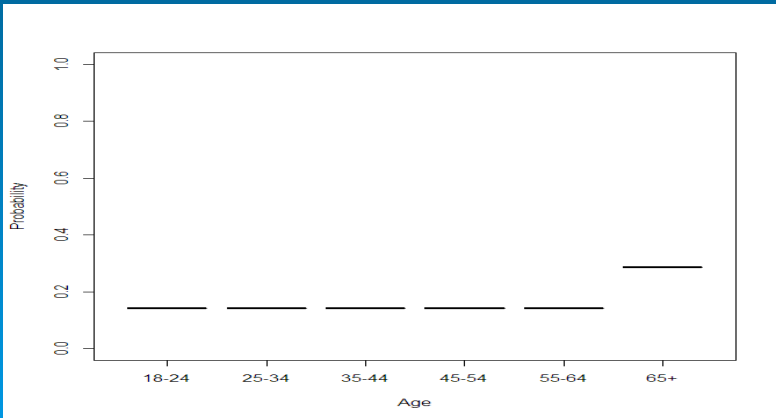




# Conclusion



## Friends

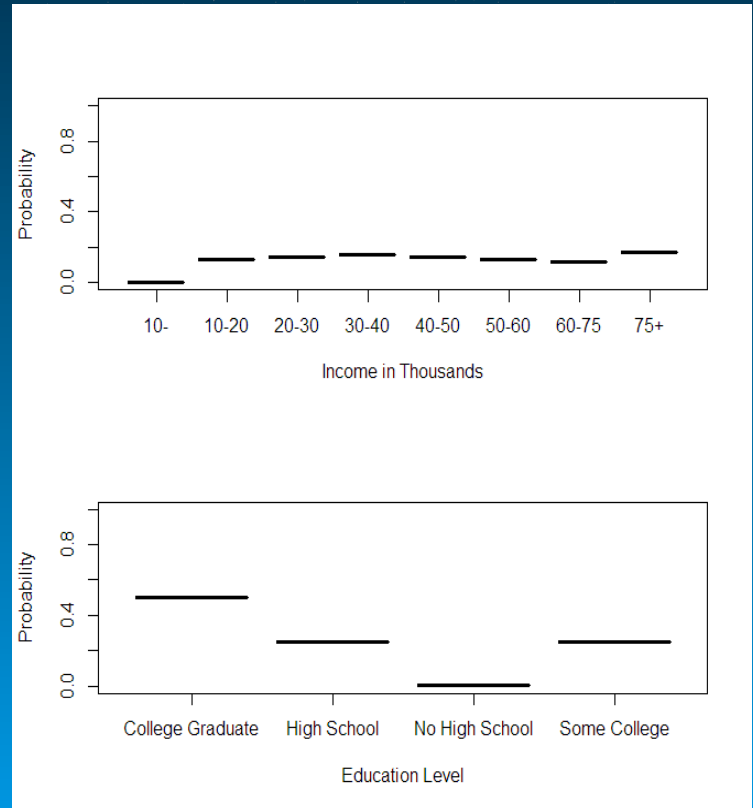
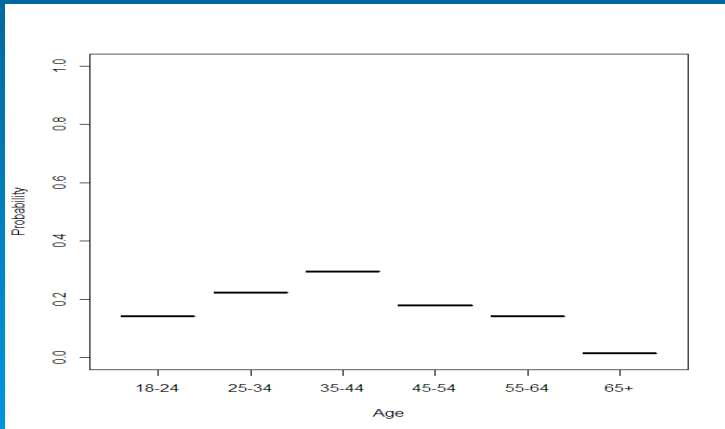




# Conclusion

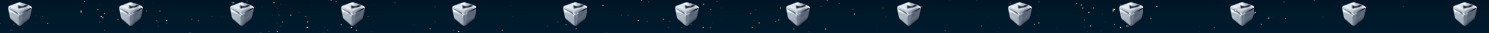


## Frasier

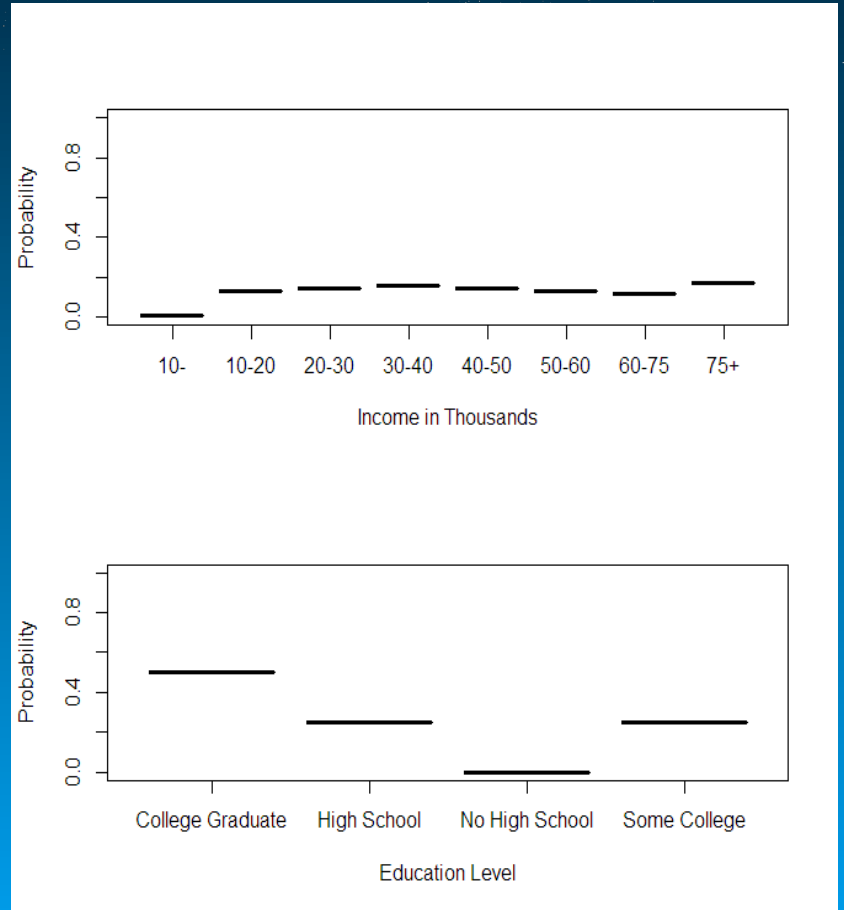
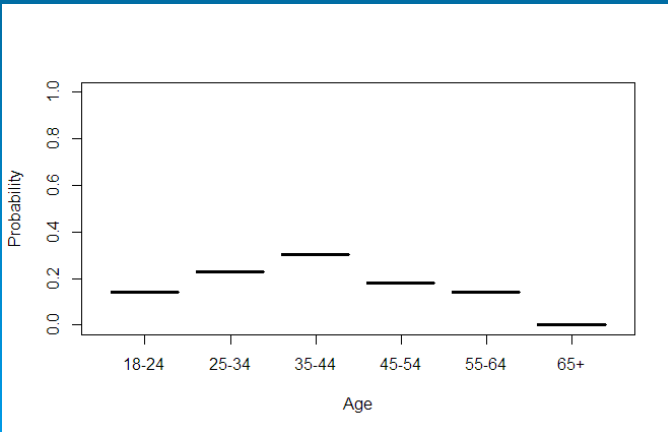




# Conclusion

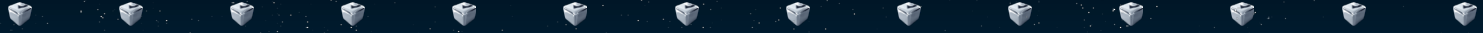


## Jesse

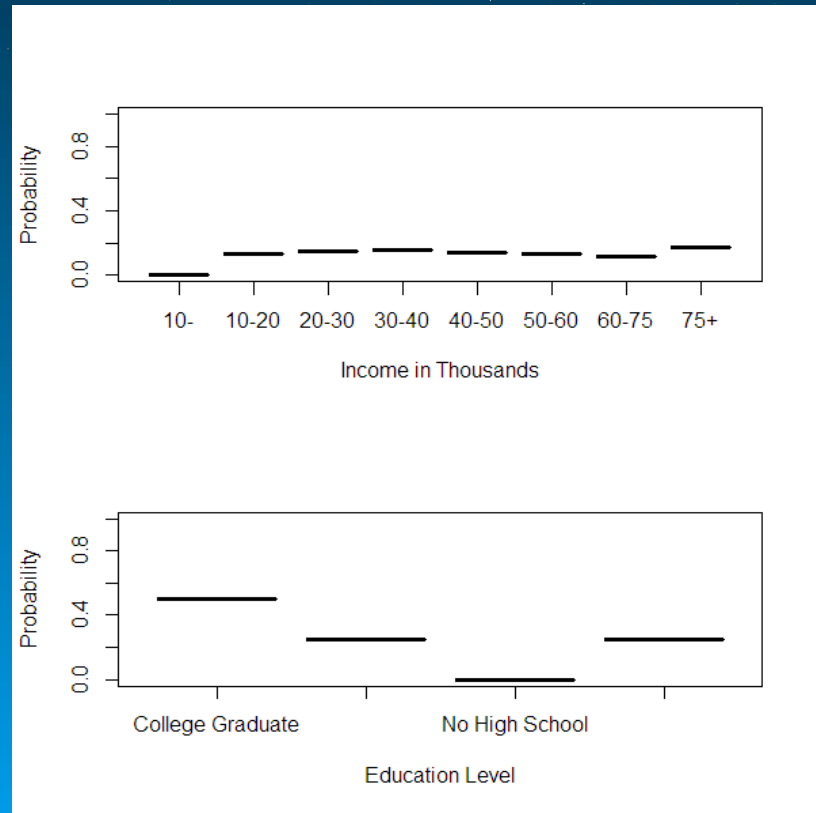
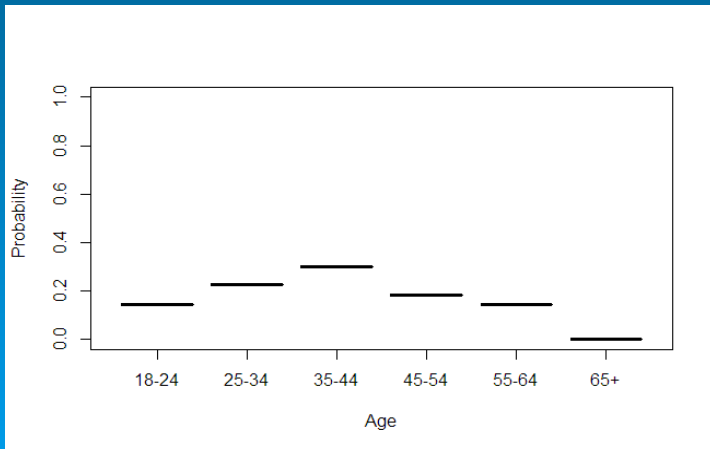




# Conclusion



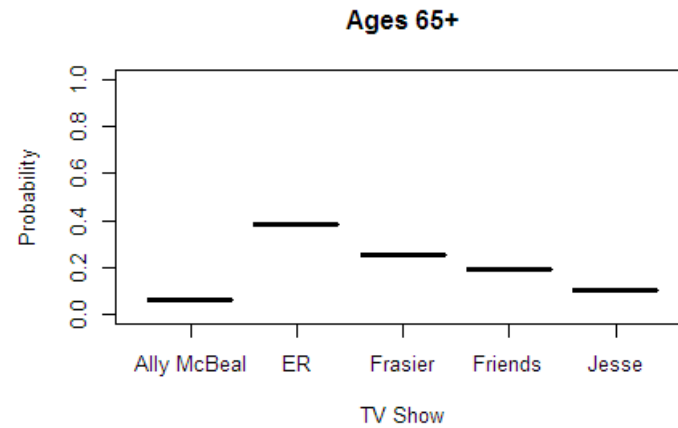
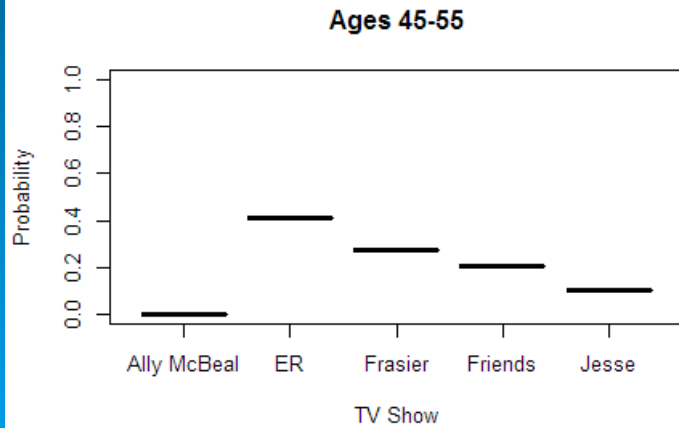
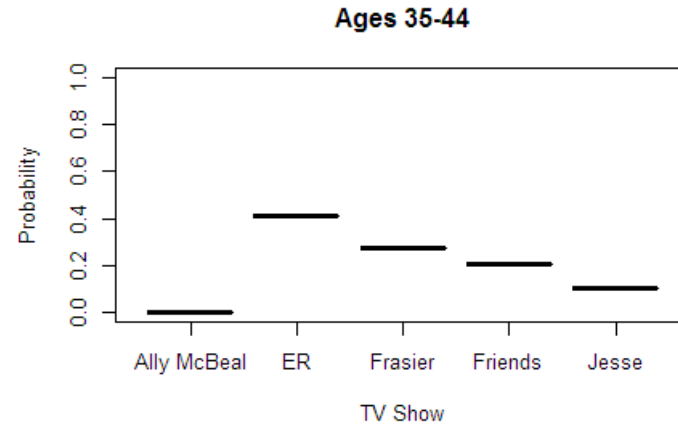
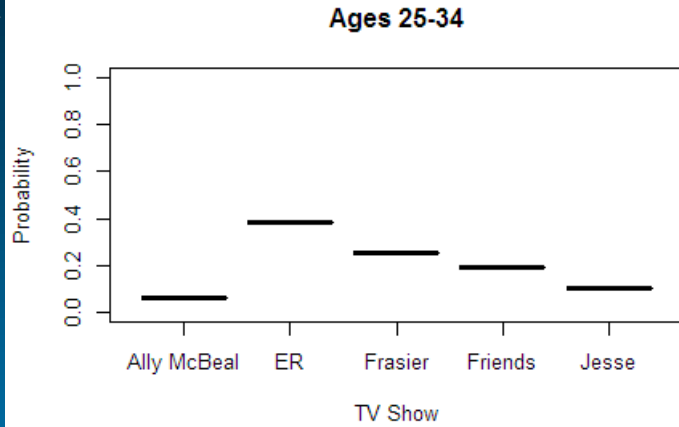
## Ally McBeal





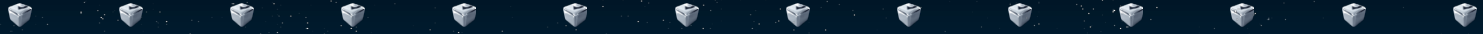
# Conclusion

## Conditional on Age

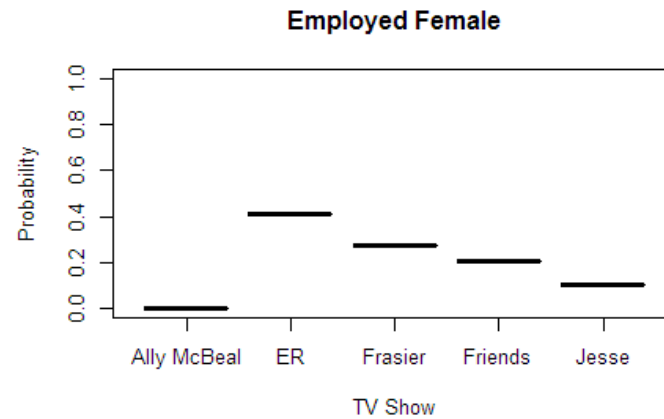
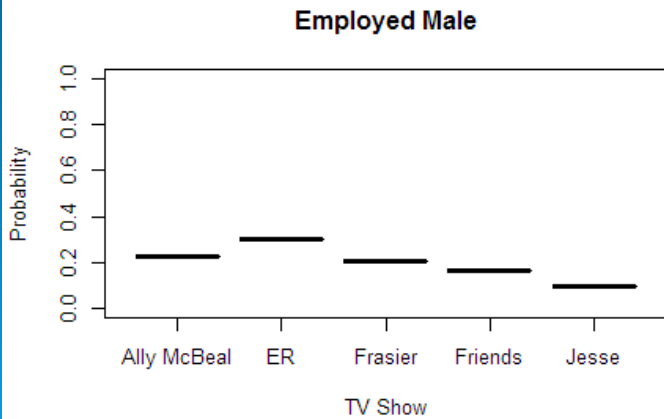
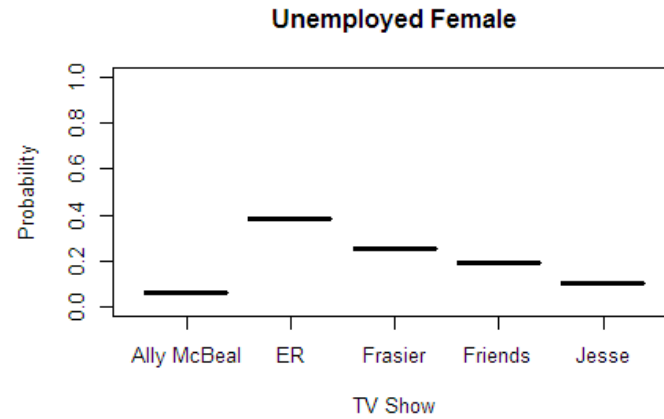
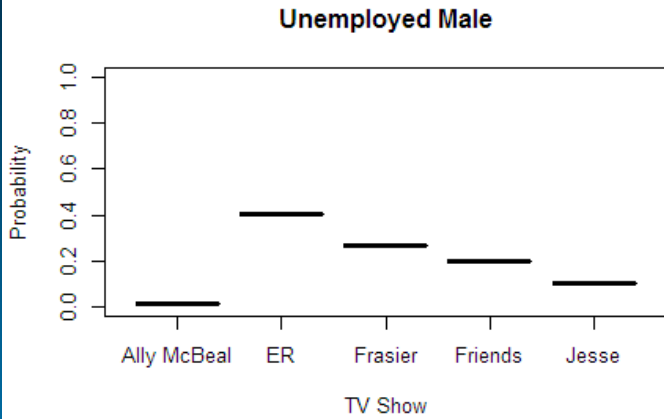




# Conclusion



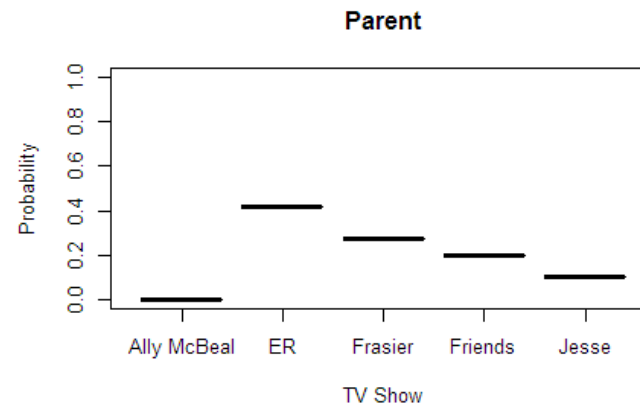
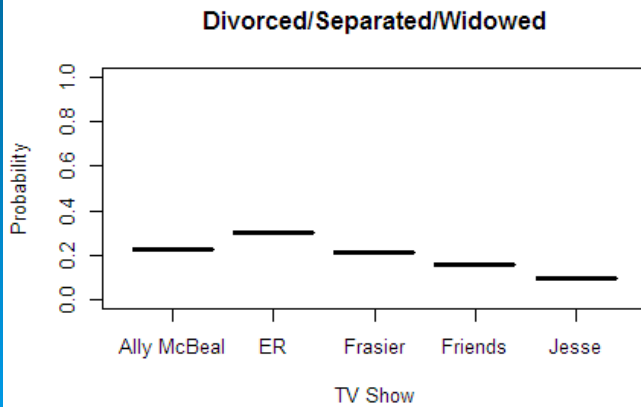
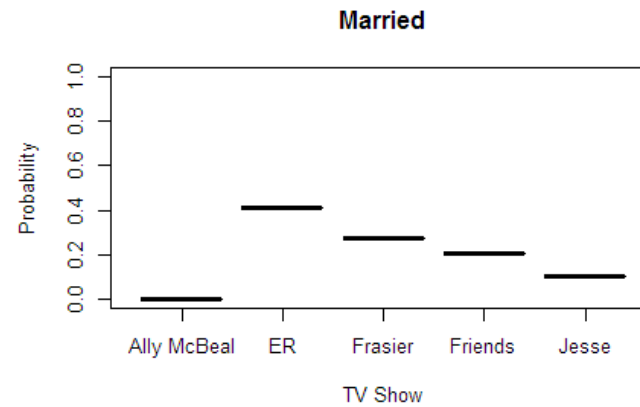
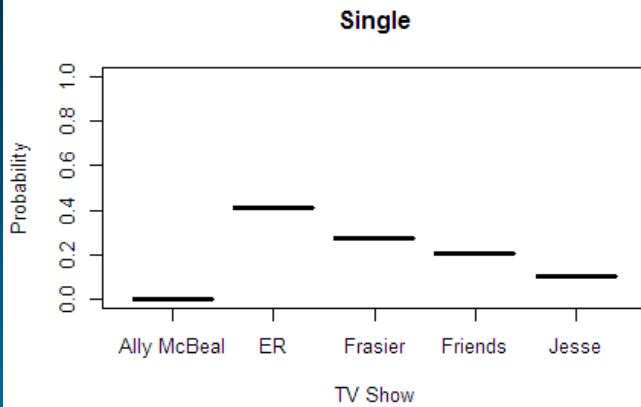
## Conditional on Employment by Gender





# Conclusion

## Conditional on Marital Status/ Parent



**Thank you**

