GENDER WAGE GAP ANALYSIS

HOANG DANG (HMD62)
GRACE LANG (GEL53)

DATA & INITIAL ANALYSIS

DATASET:

The dataset used is the Current Population Survey (CPS) taken from the Gender Pay Gap Dataset on Kaggle[1]

The current population survey is sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor statistics[2]

It contains 234 features and 344k rows from 1981 to 2013. The features detail information about a person including gender, age, employment, education, salary, etc.

1. https://www.kaggle.com/fedesoriano/gender-pay-gap-dataset
2. https://www.census.gov/programs-surveys/cps.html

DATA CLEANING:

- To make the data more usable, we removed columns with a lot of N/As. Most of them are already redundant (converted industry codes, education codes, etc.), so they can be ignored without much impact to our analysis.

- For columns that should be kept, any N/A entries were removed, which only account for < 1% of the dataset.

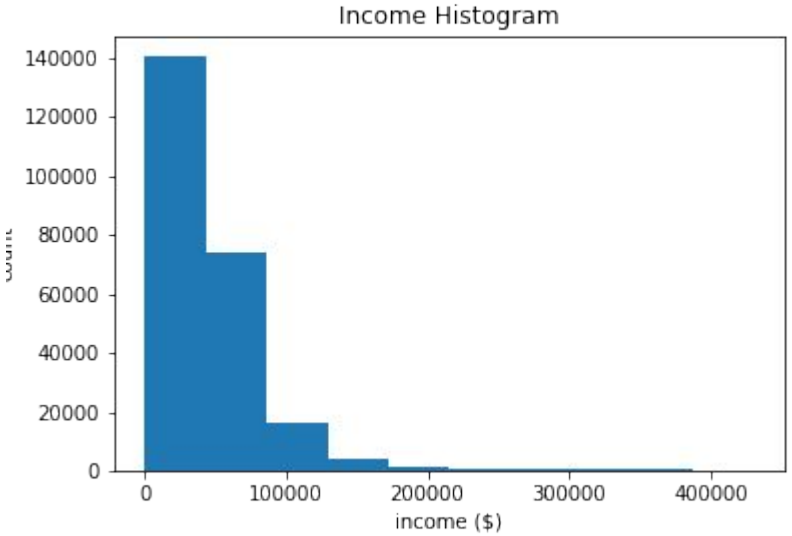
- We also removed columns with extreme class imbalance (either all 1's or 0's) because they provide no real information on how they interact with our response variables.

RESPONSE VARIABLES:

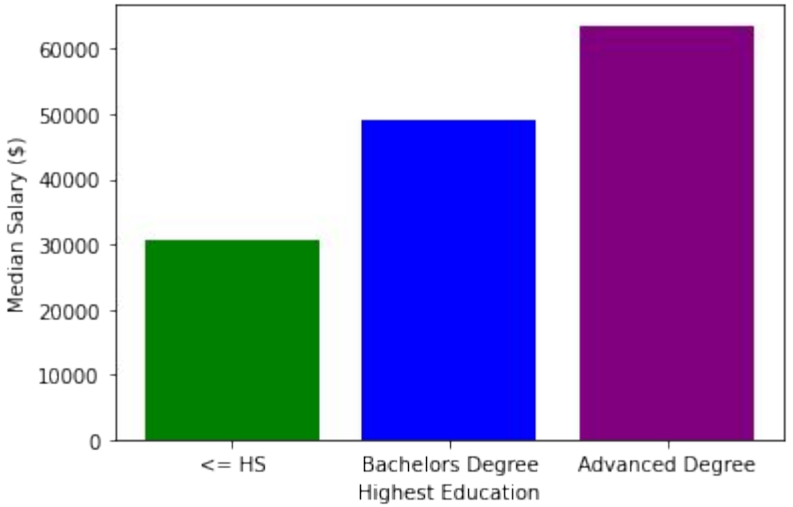
CONTINUOUS	BINARY
Annual income adjusted to 2010 dollars	Whether someone is making above the median income in 2020[3]

3. https://www.census.gov/library/publications/2021/demo/p60-273.html

DATA EXPLORATION:

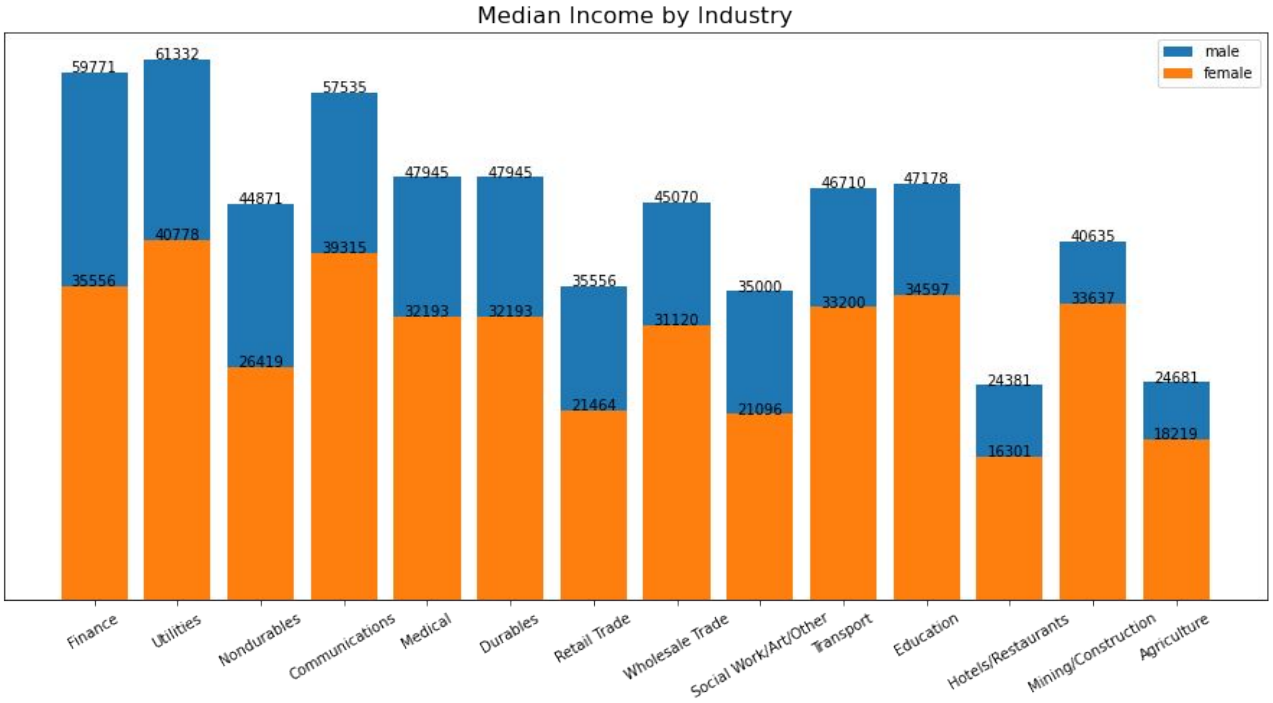


Income is right skewed



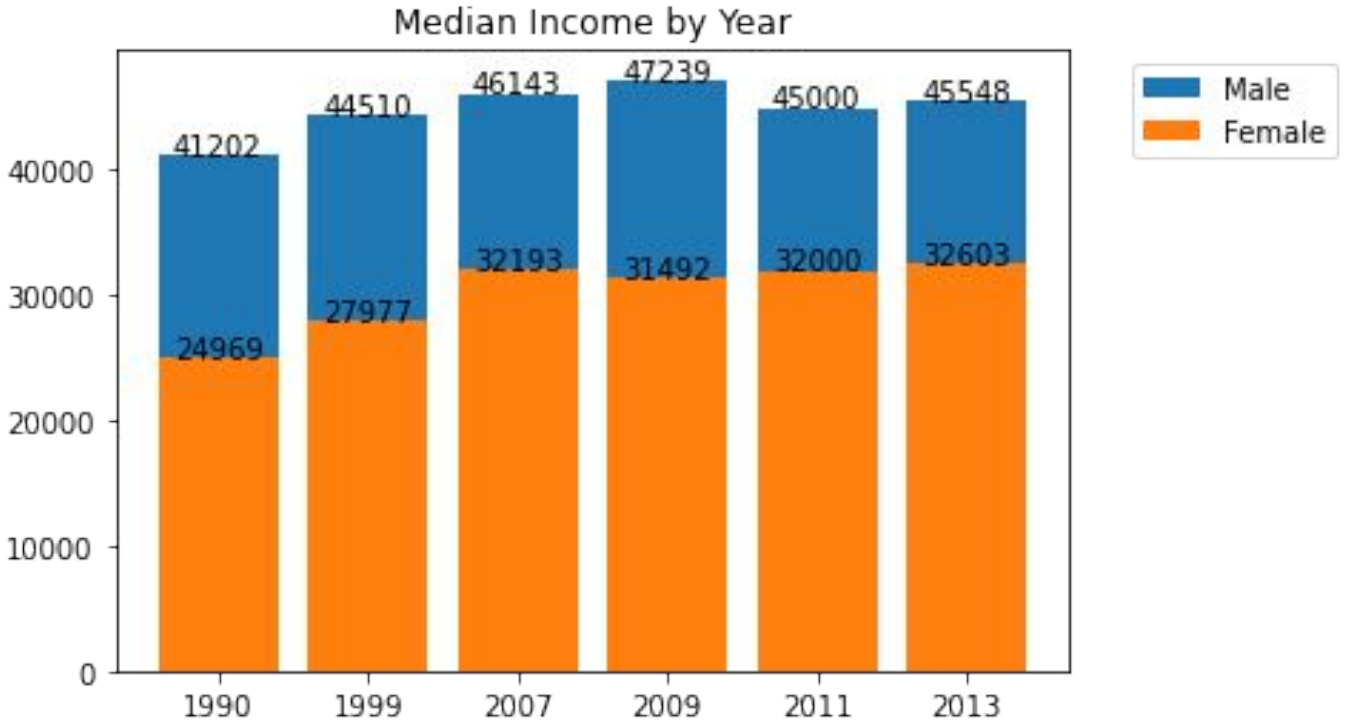
Education level is correlated with salary

DATA EXPLORATION:



Females make less than males in every industry

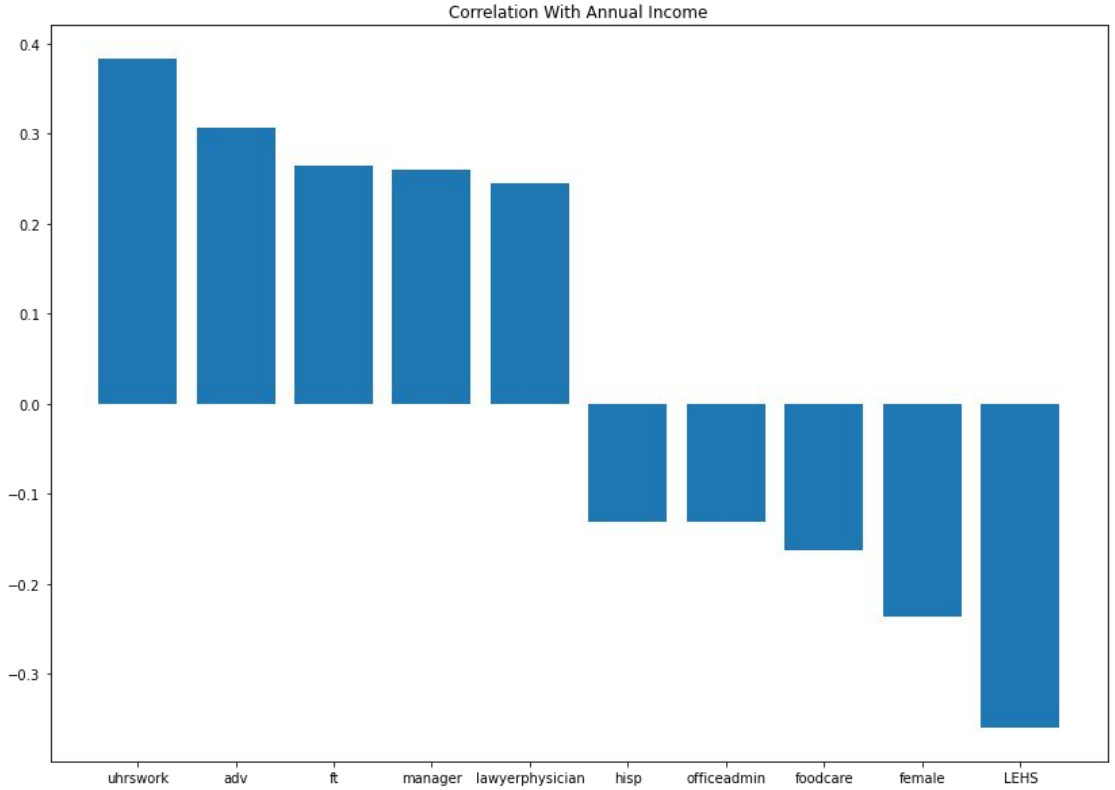
DATA EXPLORATION:



Gender wage gap is prevalent over all years, but smaller in recent time

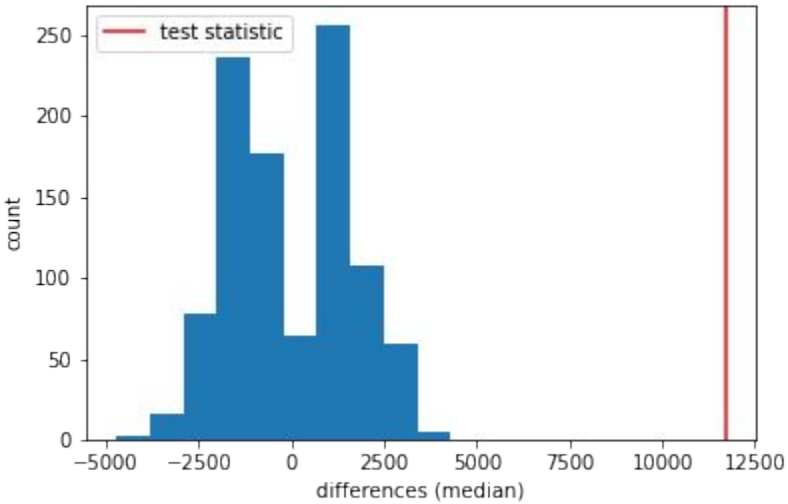
CORRELATIONS:

- Most strongly positively correlated with annual income: uhrswork, adv, ft, manager, and lawyerphysician.
- Most strongly negatively correlated with annual income: LEHS, female, foodcare, officeadmin, and hisp.



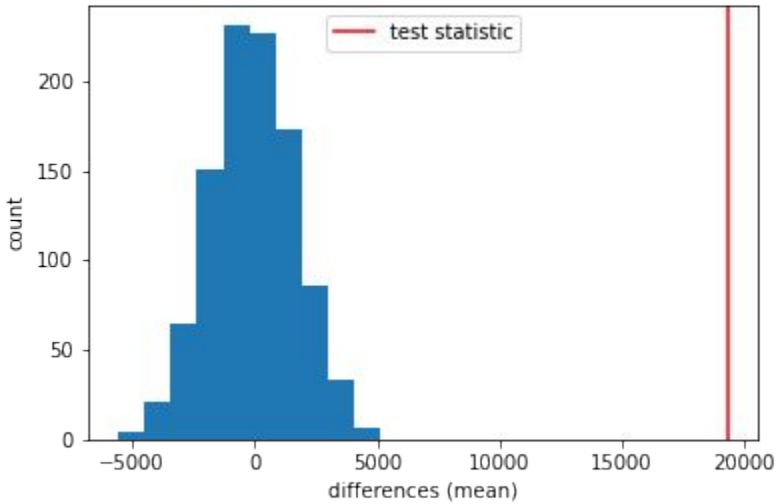
Is the income difference between genders significant?

Permutation Test: Difference in Medians



P-value = 0.0

Permutation test: Difference in Means

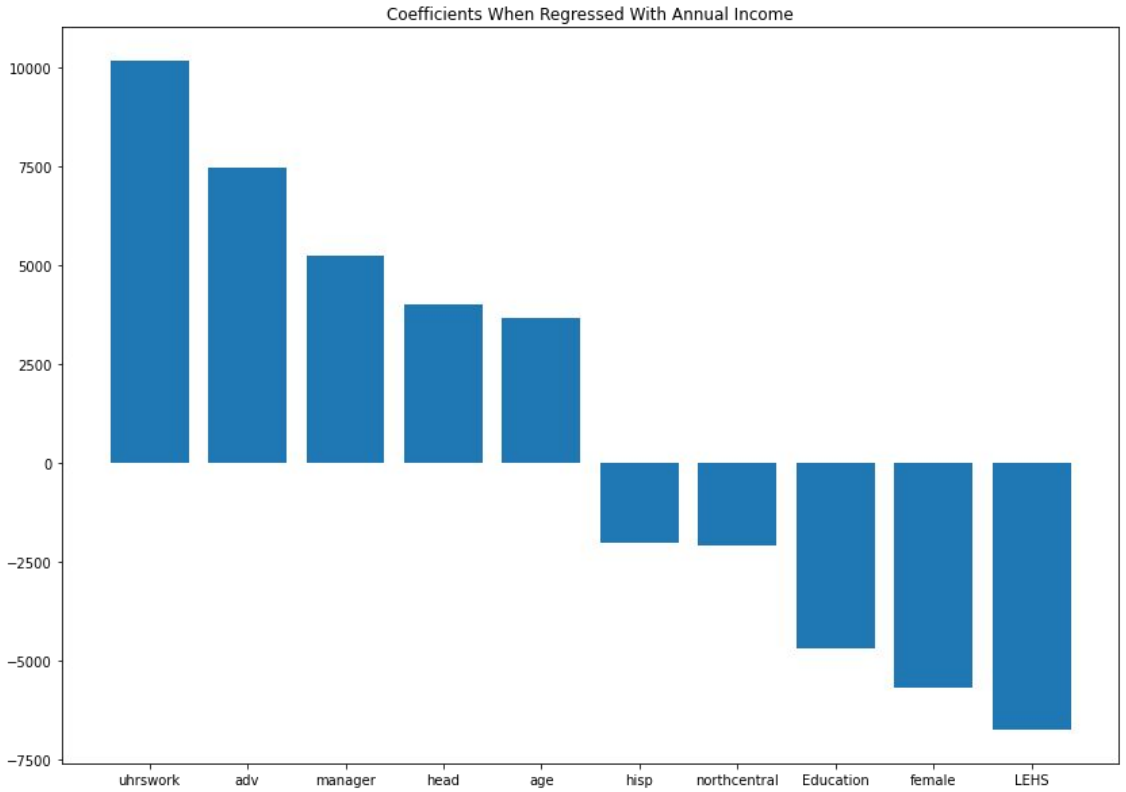


P-value = 0.0

REGRESSION ANALYSIS

LINEAR REGRESSION:

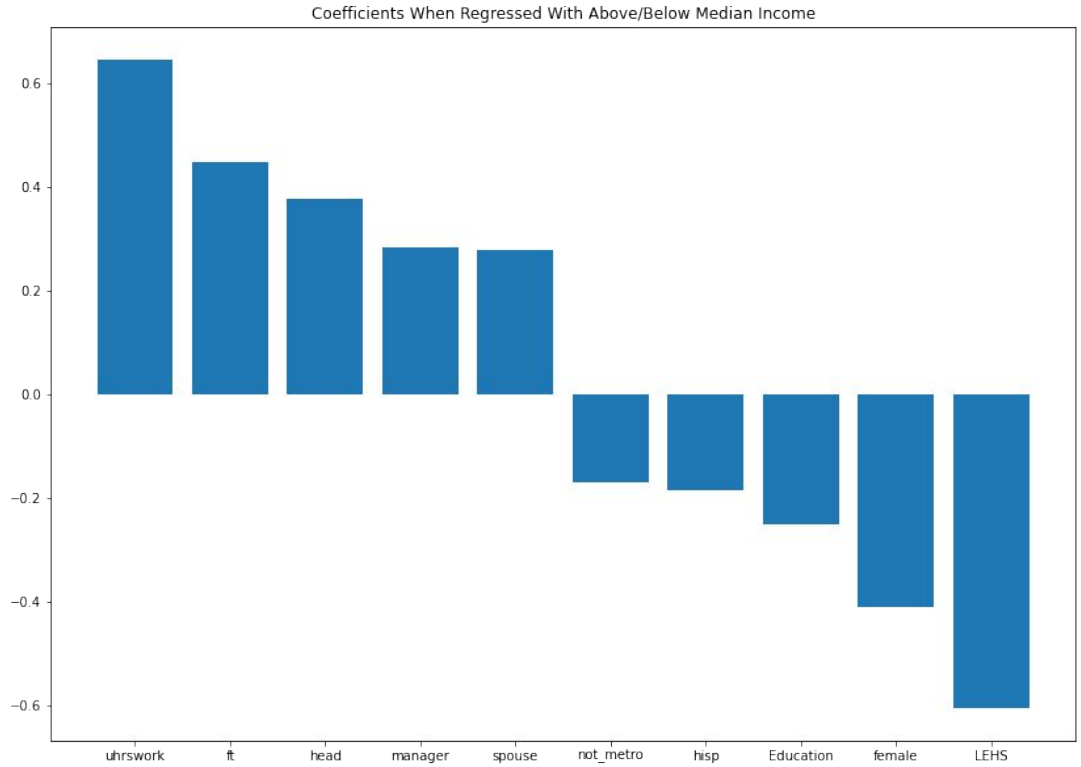
- **Strongest positive effect on annual income: uhrswork, adv, manager, head, and age ($p < 0.05$).**
- **Strongest negative effect on annual income: LEHS, female, Education, northcentral, and hisp ($p < 0.05$).**



LOGISTIC REGRESSION:

- **Strongest positive effect on whether someone makes above/below the median income: uhrswork, ft, head, manager, spouse ($p < 0.05$).**

- **Strongest negative effect on whether someone makes above/below the median income: LEHS, female, Education, hisp, and not_metro ($p < 0.05$).**



PREDICTION MODEL - CONTINUOUS:

OLS Model	MSE (CV)
With all features	985,861,721
With all features and standardized data	985,861,721
Using Lasso for feature-selection on all features	985,861,721
With all features + 2-way interactions	909,463,049
Using FSR for feature-selection on all features + 2-way interactions	921,114,815
Using Lasso for feature-selection on all features + 2-way interactions	909,011,818

PREDICTION MODEL - BINARY:

Logistic Regression Model	Total 0-1 Loss (CV)	Average 0-1 Loss (CV)
With all features	57,251	11,450.2
With all features + 2-way interactions	55,454	11,090.8
Lasso with all features	57,176	11,435.2
Lasso with all features + 2-way interactions	55,332	11,066.4
Ridge with all features	57,164	11,432.8
Ridge with all features + 2-way interactions	55,331	11,066.2

MODEL TEST ACCURACY:

OLS (CONTINUOUS)	LOGISTIC (BINARY)
MSE from predicting the mean: 1,542,374,689	Misclassification rate from predicting the most common class: 46.76%
MSE using best model: 902,771,509 (41.5% improvement)	Misclassification rate using best model: 23.27% (50.2% improvement)

CONCLUSION:

- Through the regression analysis, we controlled for certain factors like race, education, and location in order to get at the causal effects
- If ignorability holds, we saw a decrease in salary of **5253.59 corresponding to gender**
  - This would represent biased hiring practices
- However, we cannot assume ignorability, because there are idiosyncrasies not accounted for by the covariates
  - For example, there is no data on an individual's upbringing. We would like to have data on opportunity afforded to a person by their parent's education and socioeconomic status, as well as what type of school district they went to growing up

NEXT STEPS:

- **We can explore splitting income into brackets and predicting whether a person will fall into a particular bracket to improve accuracy.**
- **If time and resources permitted, we would like to explore more transformations of the features and see how they would improve model performance.**
- We would also include higher-order interactions because it could be the case that income is affected by more than 2 factors combined.
- We would also look for more recent data, and remove data from older years in order to analyze the gender wage gap solely in recent time


THANK YOU