Forecasting & Understanding Drivers Behind Job Salaries in the U.S.

Progress Report - Team 101 – GitHub Repo

ABSTRACT

The global pandemic introduced unprecedented challenges, bringing the complexities of job compensation to the forefront. This research delves into the post-pandemic changes in U.S. worker incomes, giving special attention to the interplay of demographic elements. Through a structured approach encompassing data preprocessing, in-depth exploratory analysis, hypothesis validation, and regression modeling, this study unveils subtle patterns in salary disparities.

The preliminary analysis of U.S. worker incomes reveals significant influences of demographic factors, such as gender and race, on salary disparities, with certain industries and levels of experience correlating to higher salaries. However, challenges like potential outliers, model refinement through feature selection, and exploring interactions between predictors require further investigation and sophisticated modeling approaches in subsequent phases of the research.

The data, code, and associated visualizations can be found in the Team 101 GitHub repository here. For detailed instructions on setting up and running the code, refer to the README.md file in the repository.

INTRODUCTION

The professional world has always been in flux, but recent global events have amplified the need to understand the nuances of job compensation. Deep-rooted issues of pay disparities, long a point of contention, have now become central to discussions about professional equity and justice. As job landscapes change and economic divides become more pronounced, there's an urgency to dissect the dynamics of worker compensation, especially for those in more vulnerable wage brackets [1]. This study aims to uncover the different elements that influence salary dynamics, with a keen interest in understanding the potential impact of pandemic-related disruptions on salary scales and the role of deep-rooted biases in these changes.

Our methodology is in-depth and structured. It starts with data cleaning and preparation, then moves to a comprehensive exploratory data analysis (EDA) that delves deeper into various aspects such as gender and race. Single-variable and multivariate analyzes are conducted, enhanced by statistical methods to confirm the identified patterns. Other steps in feature engineering refine the presentation of the data, such as the use of dummy variables, and improve the accuracy of the model, for example through logarithmic transformations.

Our research stands out due to its holistic view, blending both demographic and professional metrics to understand salary determinants. Rather than analyzing factors in silos, we've looked at their collective impact, especially in a post-pandemic world. By focusing on often-neglected lower-wage workers, our study provides insights that are both fresh and broadly relevant.

DATA & PREPROCESSING

First, let us understand the dataset; it comprises respondents' professional and demographic details, including age, industry, job title, annual salary and bonus, state of employment, years of experience, education level, gender, and race. Let us provide a concise walkthrough of the data preprocessing and cleaning steps:

- **Column Renaming:** We rename some columns for clarity as well as standardize them (by lowercasing them), this includes columns like “age”, “industry” and “annual_salary”. This is because the original columns are in the form of survey questions.
- **Data Type Conversion:** We convert data types so that we can treat appropriate variables as categorical; for instance, “overall_years_experience” original type is a character, and it is converted to a factor.

- **Data Standardization:** We standardize the content of certain columns; for the industry name, there are a specific number of inputs, but the survey responders were given the choice to enter a text; therefore, there is a lot of inconsistency across the same industry names. We implemented a regex pattern matching to standardize those names. For instance, any mention related to "computing" or "tech" was mapped to "Computing or Tech", "accounting", "banking", or "finance" was mapped to "Accounting, Banking & Finance", and so on. We removed the names that didn't match our patterns and managed to filter out only 4% of the dataset rows.
- **Missing Values Treatment:** Various treatment options for missing values exist, whether removing them or imputing them (i.e. MICE, mean/median/mode, neighbor techniques etc.) but we'll keep it simple for our columns that contain missing values, we filled empty values for annual_bonus with 0s, replaced empty strings of some columns (I.e. education level) with N/A so we can drop it, and also drop rows with empty "state" (since it is important for us to know which state someone is from in our study), and for race, remove the empty rows as well as rows where "optional not listed or prefer not to answer" since this variable is important to us and we can't simply put Other (since this option represents both other and potentially in an existing option).
- Since our only continuous variables are annual_salary and annual_bonus, we will simplify our analysis by adding up those two columns.

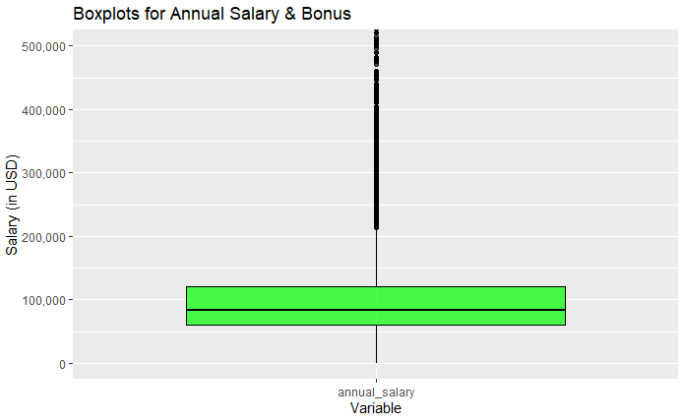
Our transformation steps above facilitate easier data handling and analysis, as well as ensuring more consistency. So now, we can move on to EDA, Feature Engineering, and Modelling.

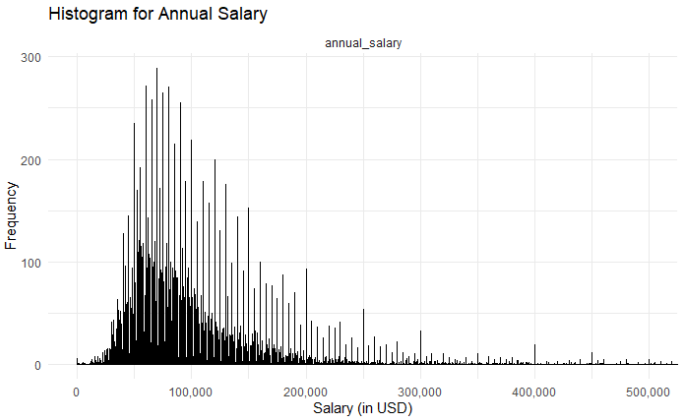
EXPLORATORY DATA ANALYSIS

EDA is the foundational step in our analysis, shedding light on underlying patterns, potential anomalies, and relationships among variables. The dataset encompasses various demographic and professional attributes of respondents, including age, industry, job title, annual salary, state of employment, years of overall and field-specific experience, education level, gender, and race. With such a multifaceted dataset, our EDA aims to unveil the nuanced factors that might influence an individual's salary. However, we do need to be careful that this dataset had voluntary respondants, and therefore suffers from a sampling bias. We cannot generalize those salaries beyond this sample due to the inherent non-randomness of this survey's sample. First, we explore the distributions of individual variables via univariate analysis, and then explore their relationships in our multivariate analysis.

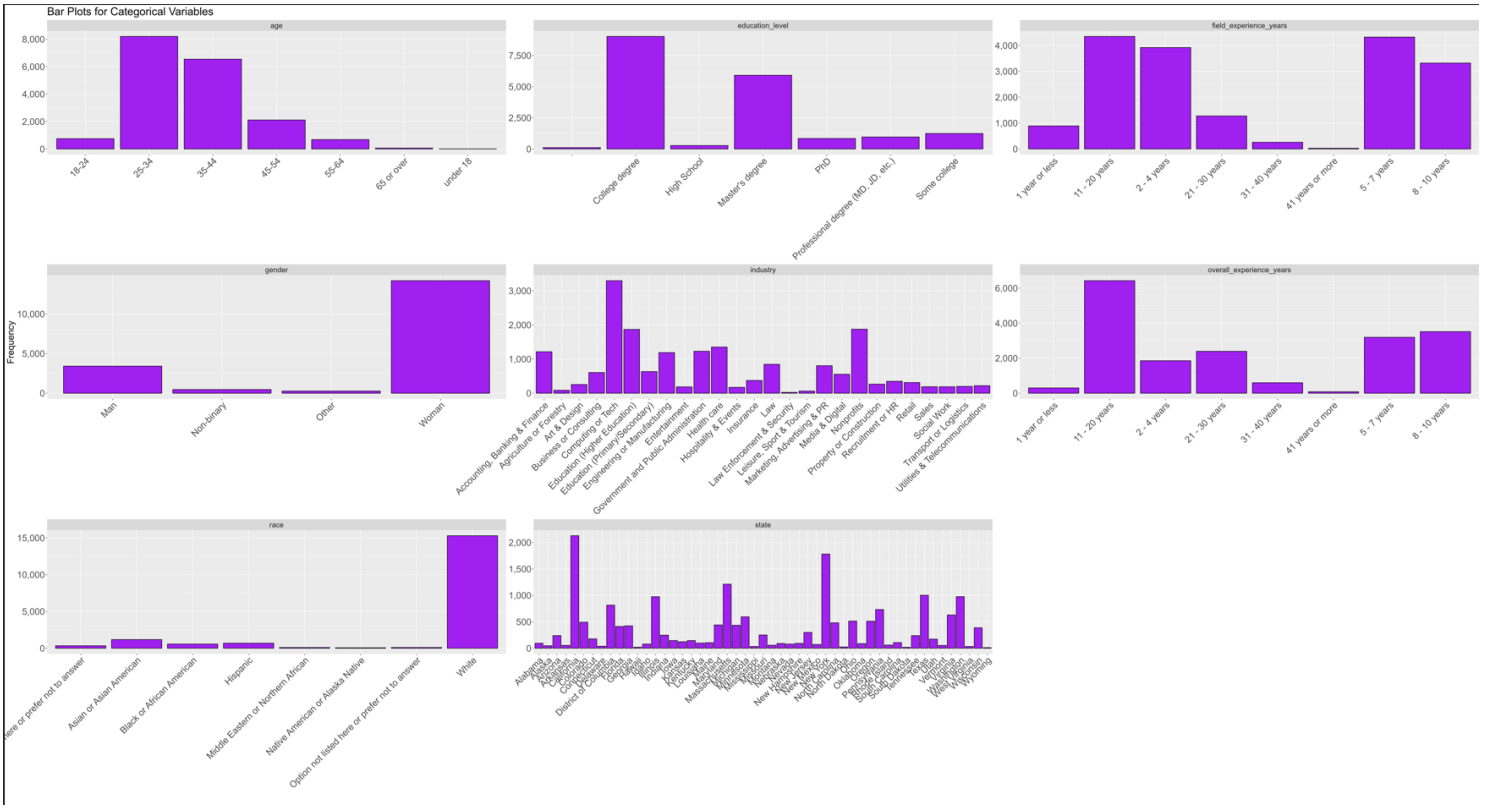
Univariate Analysis

The salary is meant to be our target variable. It is our only continuous variable in our dataset, so let us explore its distribution without considering differences across other variables. The salary ranges from USD $0 to $3 600 000, with a median of $83 000, mean of $103 524, and first quartile $59 751 and third quartile $121 000. We can visualize this with a barpot





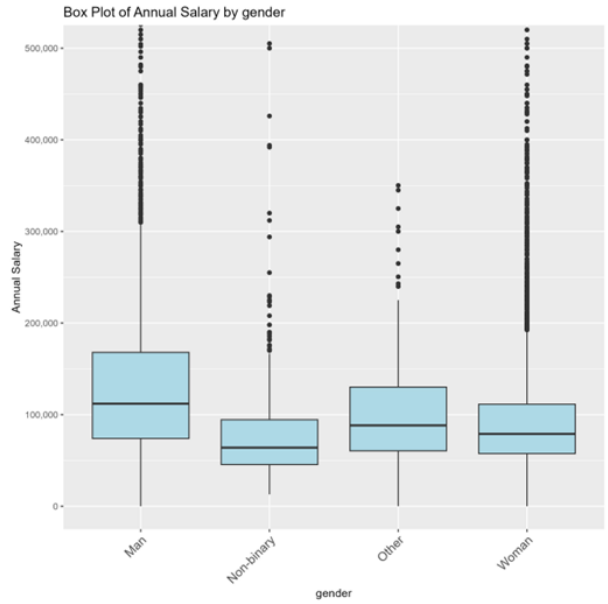
For analyzing categorical variables, let's first see the count per category. We identify that most respondents hold a college degree, followed by a master's degree. The other categories (high school, PhD, professional degree) have less than 1000 respondents each associated with them. In terms of gender, most respondents are female (14 209), followed by men (3 423). The states with the most respondents are California (2 130), followed by New York (1 783). We can observe the counts for the other categories, including industry and race, in the plots below.

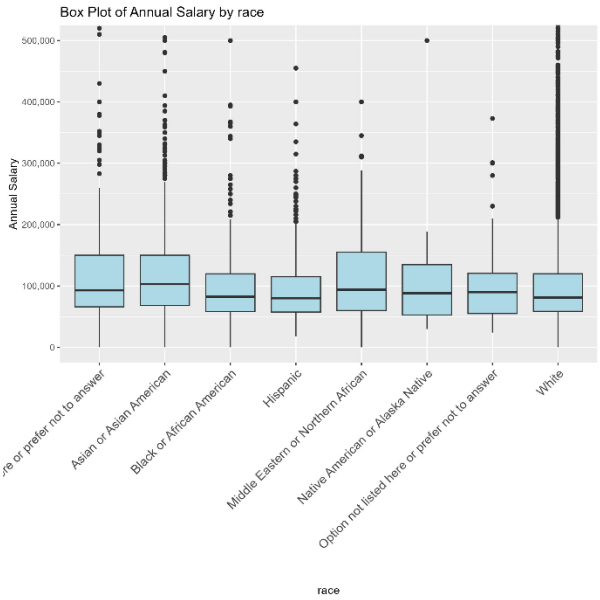


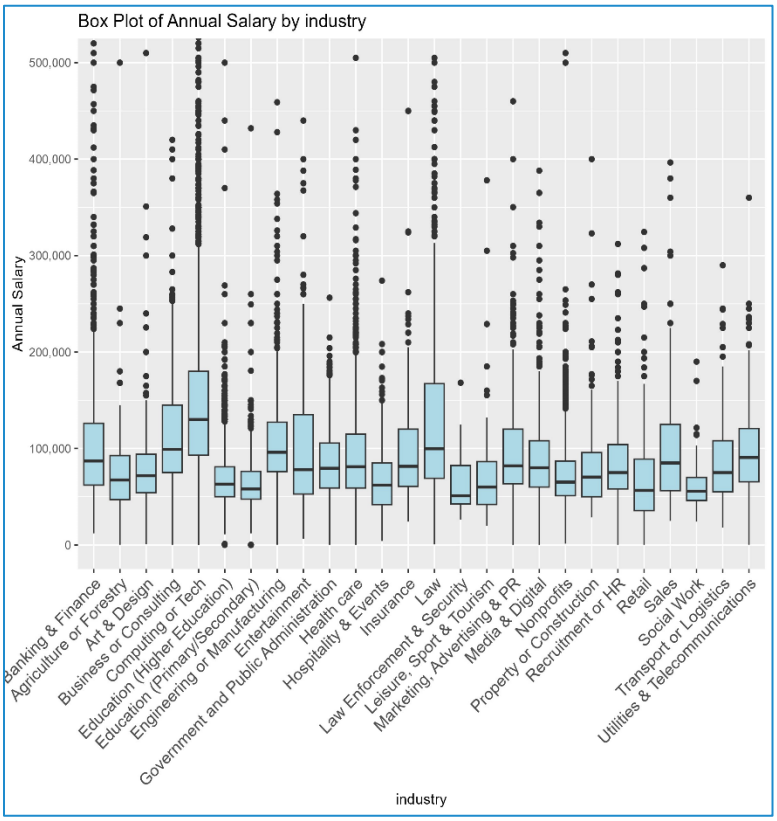
We can determine that the majority of respondents are white, women, hold a college or master's degree, are between the ages of 25 to 44, hold 11-20 years of work experiences, work in computing, education, nonprofits, and accounting/banking/finance, and reside in California or New York. This provides us with insights into the dataset's representativeness and is essential for understanding any potential biases in subsequent analyses.

Multivariate Analysis

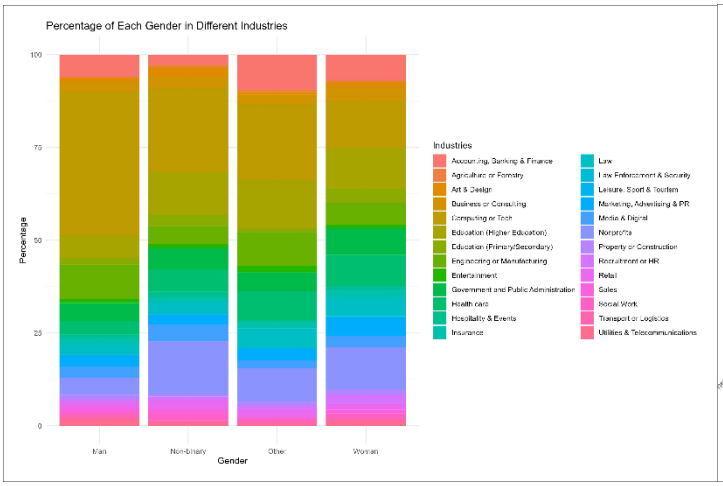
We can now analyze salaries across gender, race, and age categories revealed potential disparities. Let us first visualize the distribution of salaries across races. We identify that the race with the most variability is White. However, the race with highest median salary is Asian/Asian American, at around $100 000, followed by Middle Eastern or North African, at $94 000, while the race with lowest median salary is Hispanic, at approx. $80 000, followed by white, at $81 000. In terms of gender, we observe that men have the highest median salary, at $112 000, while women's stand at $79 000. Those insights give us a preliminary bias into salary differences amongst races and genders.

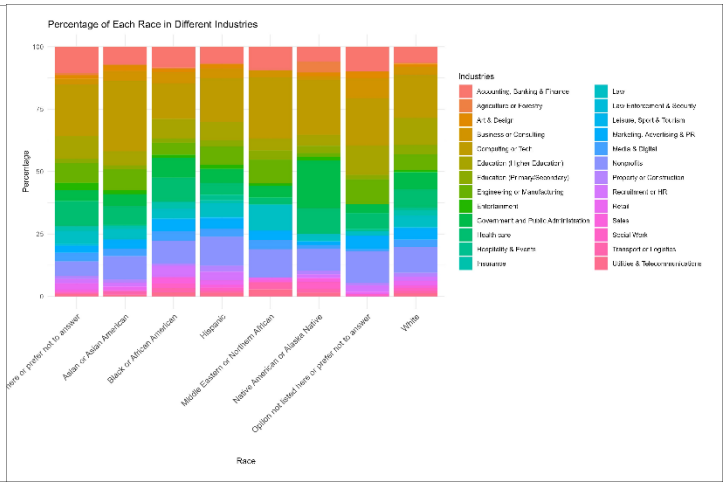




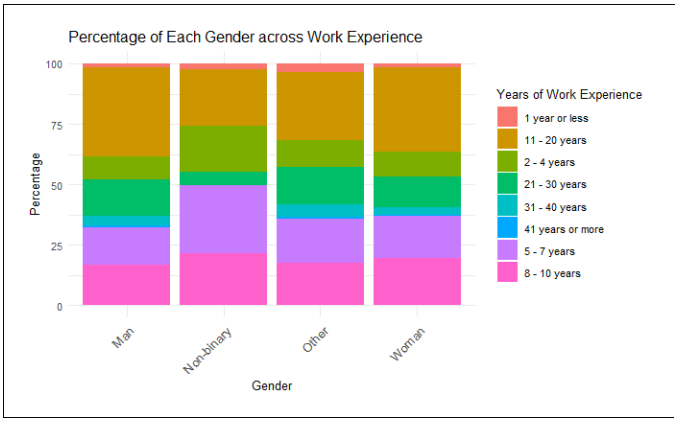


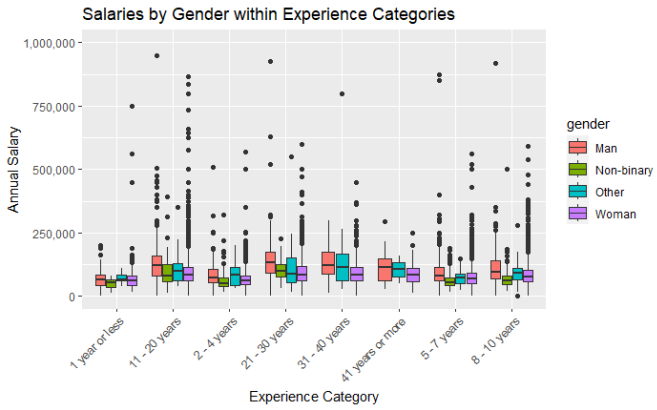
In terms of industries, we can identify that Computing or Tech pays the most (based on median), followed by Law, Business or Consulting, and Banking & Finance. We can dig deeper and find in which industries such genders and races work, to see if there might be any confounder. Since there is an imbalance of White and female responders, we'll use percentages below instead of absolute values. By visualizing the bar plot below on our right, we notice that there are no significant differences in the industry distribution across Asian Americans and Hispanics, which can make us hypothesize that there might be a bias. While there are more Hispanics working in traditionally less paying industries, such as HR and Social Work, that percentage is capped at 3%, so not significant enough. When it comes to gender, by visualizing the bar plot below on our left, we notice an overwhelming majority of men work in Computing or Tech (~ 39% of our dataset), while only third of that number of women work in that industry (~ 13%). Conversely, almost double the % of women work in Education (11% women, 6% men), triple the % of women work in Non-profits (~11.4% women, 4.6% men).





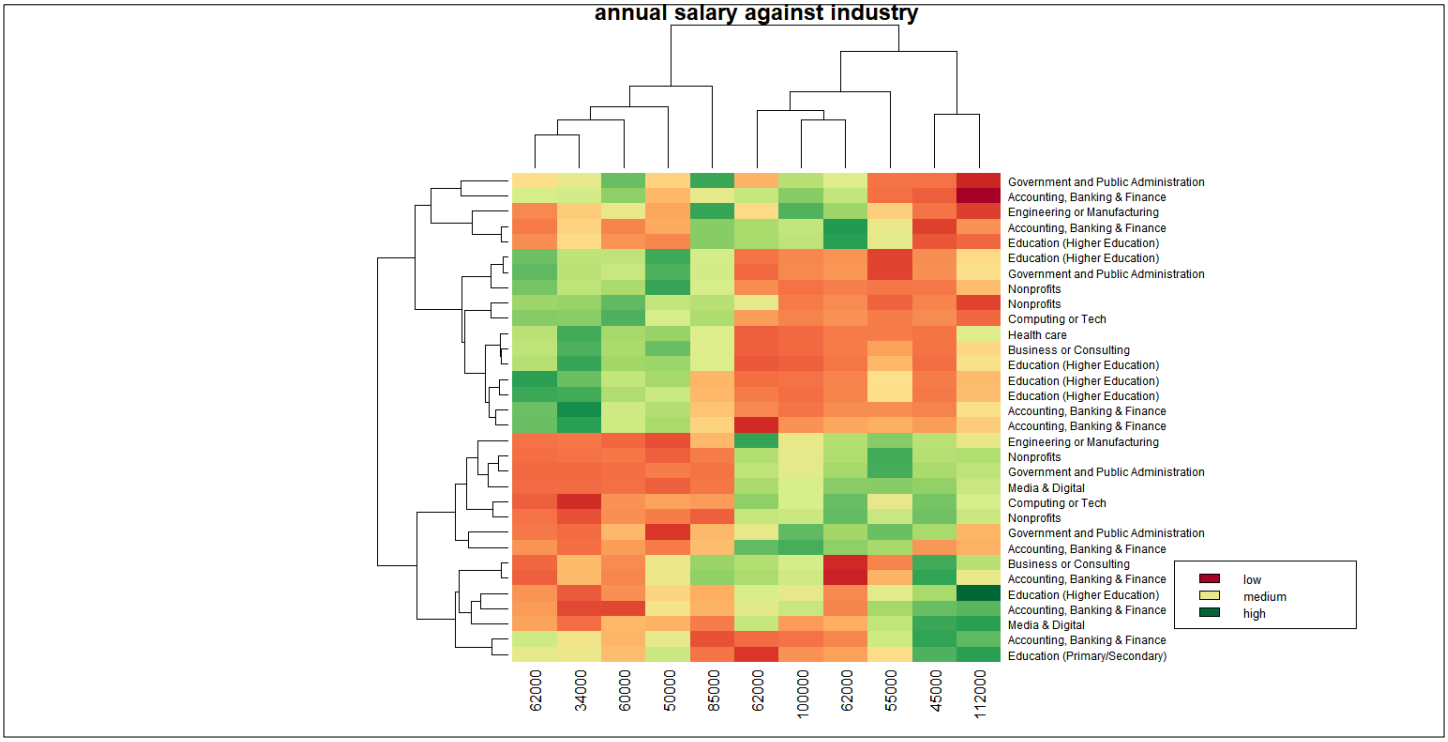
Additionally, when examining years of work experience distributions to explain pay gap between male and female, we realize the distributions are similar in our graph on the left. However, at each level of years of experience, we realize men earn (median-wise) more than women.



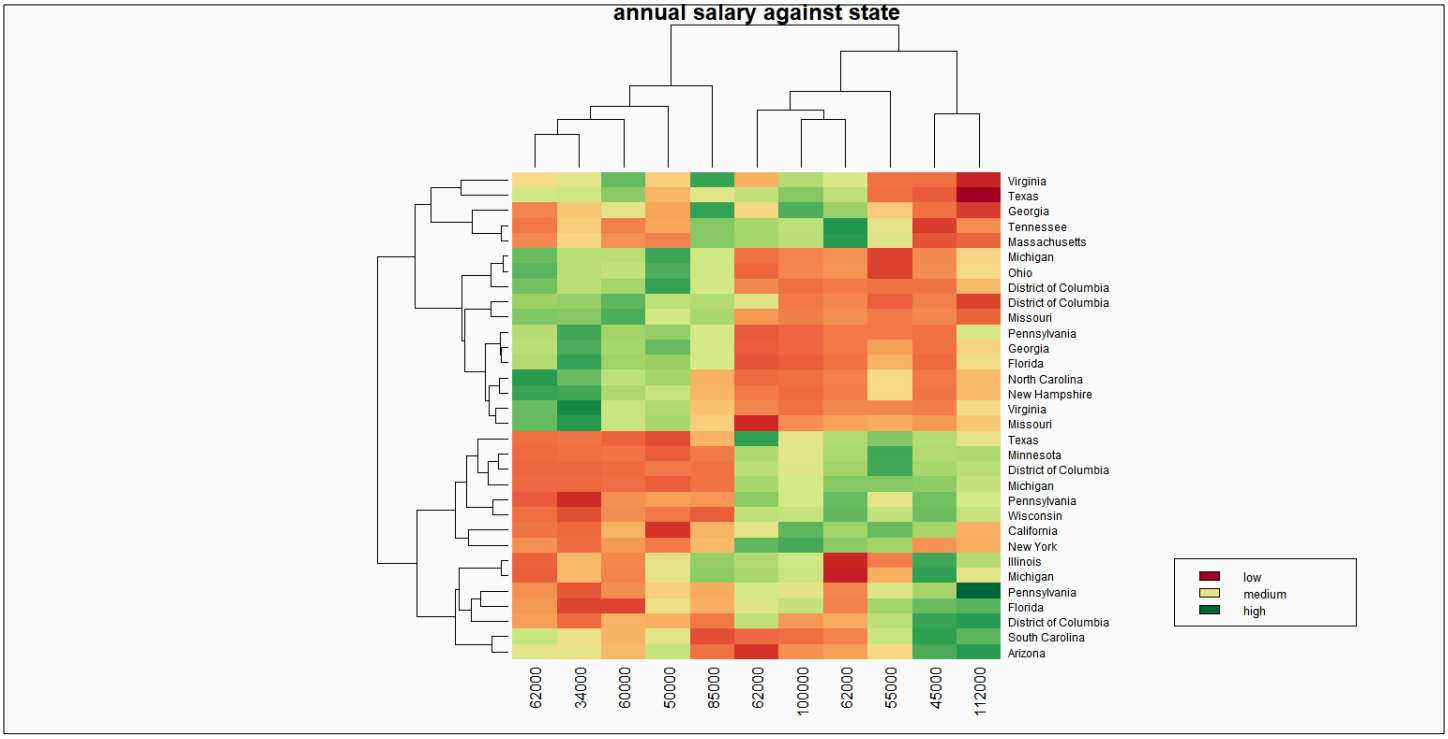


Finally, when comparing differences in races across states, we notice a large proportion of Asian Americans and Middle Eastern in a traditionally high-paying state (New York, at 15% each), while the largest proportion (16%) of Native Americans live in Oklahoma, a state with the third lower median salary, at $53 000. We can see that races do have preferences per state, and states can therefore provide an explanation as to why salaries can differ among races.

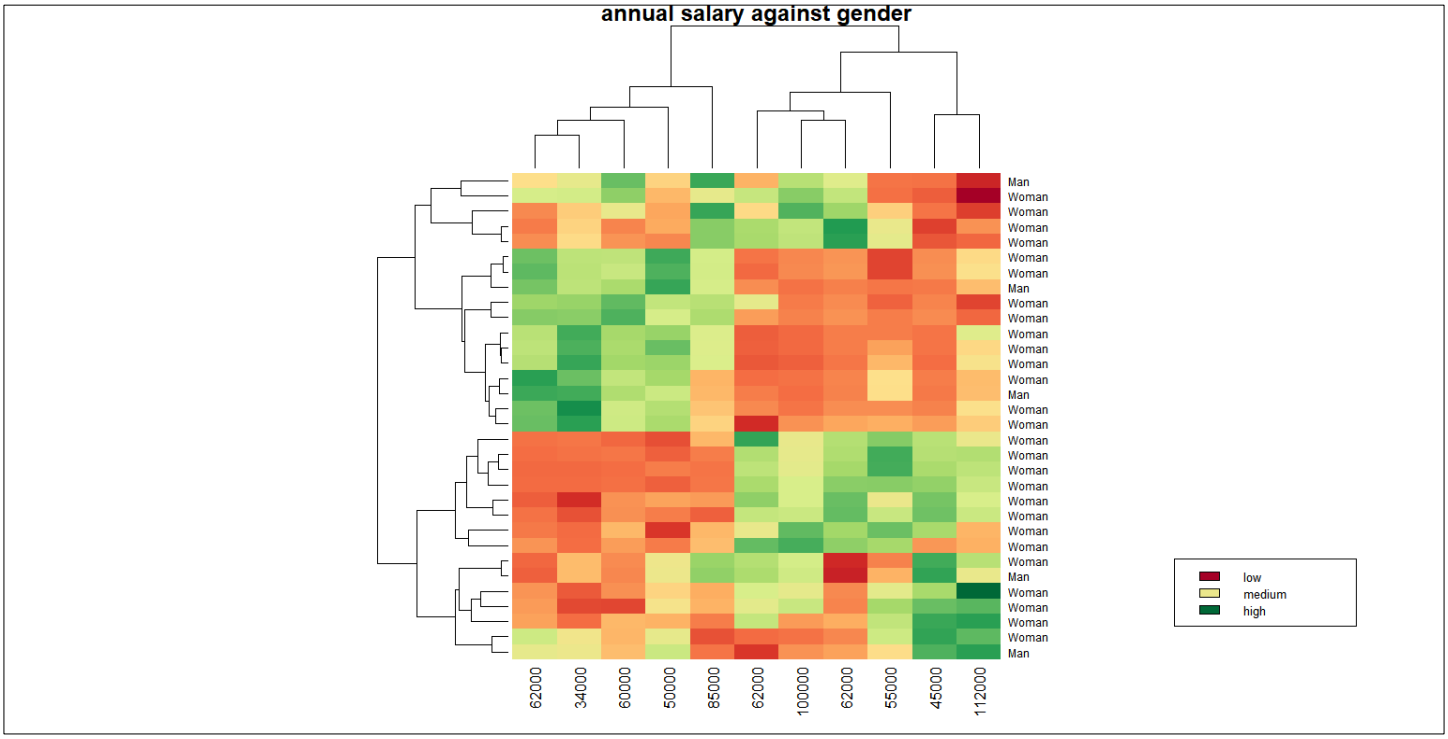
A heatmap analysis is examined for annual salary against industry, state, gender, and race with some cursory observations.



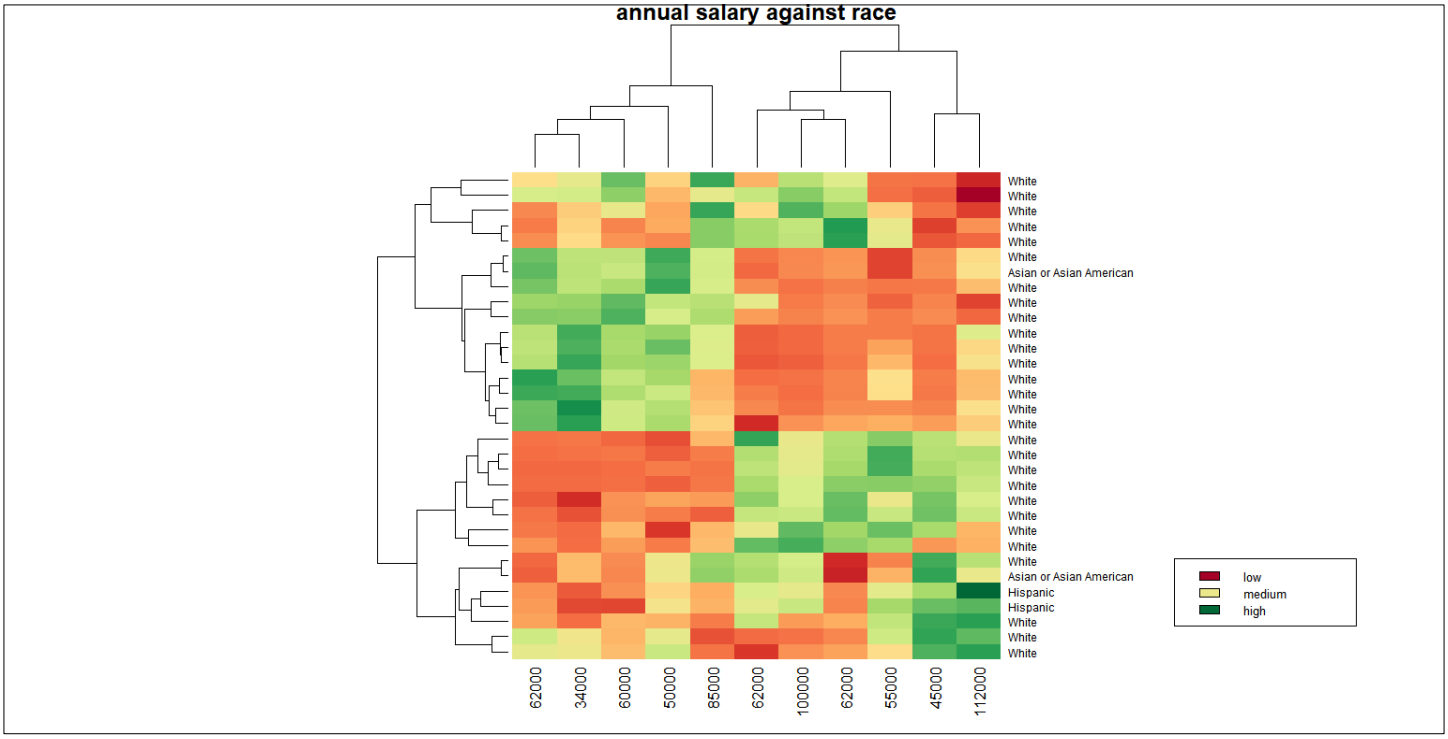
Primary/secondary educators seem to be well-paid. It is surprising because conventional wisdom suggests that primary/secondary educators are generally low-paid.



Florida and South Carolina used to be poor parts of the country. Since many foreign automobile manufacturers moved into South Carolina South Carolina prospered. Many jobs migrated to Florida and Arizona for low tax rates, favorable climates, and positive political environments. Many wealthy retirees move into Florida or Arizona.



Companies provide approximately equal compensations to men and women throughout the pay-scale range. It is surprising because conventional wisdom suggests that women tend to be low-paid for most jobs while men are typically well-paid for the same work.



Hispanics currently tend to be well-paid for high-end jobs, which white people historically used to be well-paid for.

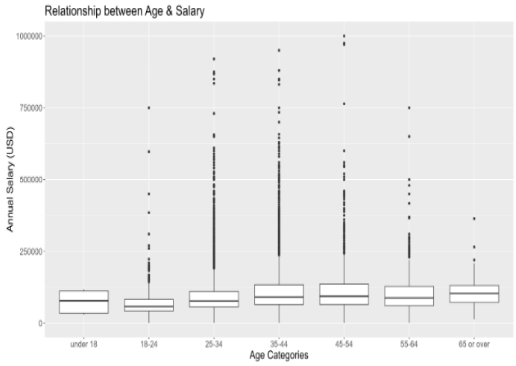
Based on what we've seen so far in our EDA, we can make some hypotheses prior to the modelling part. For instance, we can suppose that pay gaps between males and females can be explained partly by the traditionally higher % of such genders in particular domains, while gaps between races can be explained partly by their geographic preferences, which in turn impact salary. Running an ANOVA test to evaluate the effect of differences in industry, state, gender, and race, each on annual salary, give us a p-value < .05, telling us that there is a statistically significant difference amongst such attributes. Running a Chi-squared test across combinations of such attributes also give us a p-value <.05, telling us that there are differences between classes.

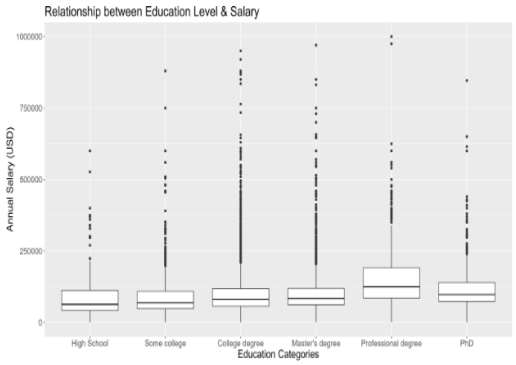
Now that we have stated some of our hypotheses, we will however further investigate by running more exhaustive modelling below.

FEATURE ENGINEERING & MODELING

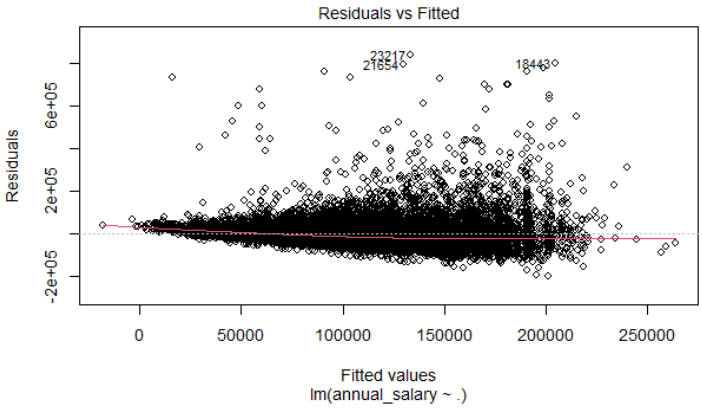
In this section, we discuss the critical steps undertaken for feature engineering and modeling choices we made. We highlight the challenges posed by categorical data, our approaches to address them, and the insights derived from our initial models.

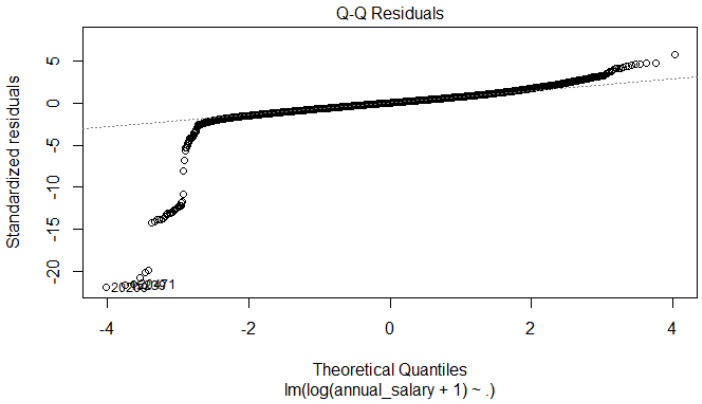
Handling Categorical Data





Our dataset comprises several categorical variables, posing the potential risk of the curse of dimensionality if not managed appropriately. We took the following steps to handle this: in terms of geographical data, the dataset contains 51 unique US states. To manage this granularity, we grouped the states into three broad regions; West, Central, and East, for states in the western, central, and eastern parts of the country, respectively. Additionally, to streamline our analysis and focus on the primary gender categories, we filtered out non-binary and other gender classifications, retaining only 'male' and 'female' categories. Our predictors are exclusively categorical, which influenced our modeling decisions. For binary predictors, like gender, the linearity assumption is inherently satisfied. In terms of nominal categorical predictors with multiple categories, such as age or education level, there's a discernible trend in the data as illustrated in the accompanying graphs. However, it's essential to approach ordinality with caution. For instance, while a Master's degree is academically a step above a College degree, the median salary doesn't reflect a substantial difference between the two. Similarly, age presents an intriguing pattern: post the 35-44 age bracket, not only does the number of upper-tail outliers decrease, but average salaries also witness a decline.





Modelling and Assumptions

Our initial model using linear regression without considering ordinality resulted in an adjusted R-squared of 0.3009. A subsequent regression with ordinality yielded similar performance, but coefficients become harder to interpret since they succumb to quadratic and cubic transformations. A log transformation on the response variable showed a minor improvement in adjusted R-squared (0.3056) and a better residual distribution. The summary table can be found in the appendix. The residuals shown on the Residuals vs Fitted plot (top right) indicate that our model might be underestimating salaries, especially at higher ranges, which could be due to right-tailed outliers. Using a Cook's distance cutoff of 4/n, where n is the number of data points (17 204), we identify 505 influential records (or 2.9% of our dataset). This will be addressed in a future iteration. However, challenges remain as the standardized residuals do not yet follow a normal distribution (Q-Q Residuals Plot under Residuals vs Fitted plot). Notably, other advanced regression techniques might offer robustness against outliers and capture non-linearity more effectively, as highlighted in the referenced paper in [2].

Key Insights from Regression Analysis

Table Showing % Change in Annual Salary Relative to Base Case

Predictor	Category	% Δ Ann. Salary
Age	25-34	10.06%
Industry	Agriculture or Forestry	-30.83%
	Art & Design	-20.6%
	Business or Consulting	11.01%
	Computing or Tech	33.52%
	Hospitality & Events	-28.36%
	Education (Lower Education)	-43.54%
	Social Work	-37.37%
Field Experience	31 - 40 years	104.1%
Region	East	13.61%
	West	19.10%
Gender	Woman	-12.93%
Race	Middle Eastern or North African	-16.71%
Education Level	Professional degree (JD, MD...)	65.34%
	PhD	41.75%

Statistically Significant Predictors: Gender, regions (Central, East, West), race, education level, and field-specific experience all demonstrated statistical significance with p-values < 0.05. Age's significance varied across its brackets, while the 'overall_experience_years' variable showed no significant distinction, indicating its potential removal in future model iterations.

Coefficients interpretation: The model's intercept was statistically significant, underscoring the importance of reference cases for our categorical variables. When predictors were at their default reference levels, the base salary was approximated at $70,601. Factors such as industry type and education level significantly influenced this base salary. To illustrate, our intercept corresponds to specific base cases like ages 18-24, years of experience (overall and field) being 1 year or less, holding a College degree, being an Asian or Asian American male, in the Central region and in the Accounting, Banking & Finance industry. Any variations from this baseline can be interpreted using the rightmost column of our table on the left. For instance, a category with a +10% change implies a 10% salary increase over the base, resulting in an average salary of approximately $77,661.

CONCLUSION

Our preliminary analysis has provided several insights into the factors influencing salaries. Demographic factors such as gender or race were shown, both in exploratory analysis and regression modelling, to play a role in determining salaries, thus indicating certain biases. As expected, certain industries and professionals with certain years of experience under their belt are proven to have higher salaries.

However, challenges persist:

- The presence of potential outliers needs addressing, possibly through imputation or exclusion, or the use of non-linear models.
- We will explore the impact of using regions vs. states and consider introducing job roles as a predictor.
- Feature selection techniques, like VIF and p-value of coefficients, will be employed to refine our model.
- Exploring interactions between categorical predictors, especially between industries and regions, will be crucial.

Future work will involve more sophisticated model selection and hyper-parameter optimization. Given that our predictors are all categorical, our current approach aligns with ANOVA. This restricts certain modeling freedoms and necessitates careful consideration of interactions between variables.

The upcoming challenges include managing the curse of dimensionality that will come with such feature considerations.

REFERENCES

[1] Johnson, Elizabeth R., and Ashley V. Whillans. "The Impact of the COVID-19 Pandemic on the Satisfaction of Workers in Low-Wage Jobs." Harvard Business School Working Paper, No. 23-001, July 2022.

[2] Kibekbaev A. and Duman E. “Benchmarking Regression Algorithms for Income Prediction Modeling”. 2015 International Conference on Computational Science and Computational Intelligence

APPENDIX

Regression Summary for our Log-Linear Model

```
Residuals:
    Min      1Q  Median      3Q     Max
-199901  -29983   -7126   16708  836582

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        98269.3     4610.2  21.316  < 2e-16 ***
age25-34                                            5026.7     2836.5   1.772 0.076380 .
age35-44                                            2392.5     3175.9   0.753 0.451253
age45-54                                           -4798.4     3651.1  -1.314 0.188787
age55-64                                          -12097.4     4611.6  -2.623 0.008717 **
age65 or over                                      -7825.9     9129.0  -0.857 0.391316
ageunder 18                                        13708.9    24832.6   0.552 0.580921
industryAgriculture or Forestry                   -30042.5     6957.1  -4.318 1.58e-05 ***
industryArt & Design                              -26588.7     4330.6  -6.140 8.45e-10 ***
industryBusiness or Consulting                      8451.4     3090.0   2.735 0.006243 **
industryComputing or Tech                          34425.0     2121.3  16.228  < 2e-16 ***
industryEducation (Higher Education)              -48053.2     2354.8 -20.406  < 2e-16 ***
industryEducation (Primary/Secondary)             -49520.8     3046.3 -16.256  < 2e-16 ***
industryEngineering or Manufacturing               -4703.7     2548.3  -1.846 0.064935 .
industryEntertainment                              -6292.0     4983.6  -1.263 0.206774
industryGovernment and Public Administration      -32964.1     2526.6 -13.047  < 2e-16 ***
industryHealth care                               -17780.1     2468.3  -7.203 6.12e-13 ***
industryHospitality & Events                      -30307.1     5045.4  -6.007 1.93e-09 ***
industryInsurance                                  -8659.2     3646.9  -2.374 0.017589 *
industryLaw                                        -7453.7     3147.3  -2.368 0.017882 *
industryLaw Enforcement & Security                -33708.7    12944.3  -2.604 0.009219 **
industryLeisure, Sport & Tourism                  -33241.0     7848.3  -4.235 2.29e-05 ***
industryMarketing, Advertising & PR                -7299.9     2812.9  -2.595 0.009463 **
industryMedia & Digital                           -18610.6     3207.9  -5.802 6.69e-09 ***
industryNonprofits                                -36472.9     2304.2 -15.829  < 2e-16 ***
industryProperty or Construction                  -21731.2     4183.6  -5.194 2.08e-07 ***
industryRecruitment or HR                         -11230.5     3747.0  -2.997 0.002729 **
industryRetail                                    -32507.3     3983.0  -8.162 3.53e-16 ***
industrySales                                       1101.5     4815.7   0.229 0.819081
industrySocial Work                               -45295.5     4888.6  -9.266  < 2e-16 ***
industryTransport or Logistics                    -14076.2     4678.7  -3.009 0.002629 **
industryUtilities & Telecommunications             -8955.2     4483.7  -1.997 0.045813 *
overall_experience_years11 - 20 years                783.7     4869.4   0.161 0.872138
overall_experience_years2 - 4 years                 -9835.1     4544.3  -2.164 0.030457 *
overall_experience_years21 - 30 years               -2491.6     5250.7  -0.475 0.635130
overall_experience_years31 - 40 years               -6676.3     6282.6  -1.063 0.287949
overall_experience_years41 years or more            -8487.1     9703.5  -0.875 0.381783
overall_experience_years5 - 7 years                 -6982.6     4713.7  -1.481 0.138536
overall_experience_years8 - 10 years                -2494.3     4765.3  -0.523 0.600686
field_experience_years11 - 20 years                 48991.9     2983.6  16.420  < 2e-16 ***
field_experience_years2 - 4 years                   10928.3     2733.8   3.997 6.43e-05 ***
field_experience_years21 - 30 years                 69704.4     3623.7  19.236  < 2e-16 ***
field_experience_years31 - 40 years                 73329.2     5642.7  12.995  < 2e-16 ***
field_experience_years41 years or more              48972.7    13152.6   3.723 0.000197 ***
field_experience_years5 - 7 years                   22738.1     2837.0   8.015 1.17e-15 ***
field_experience_years8 - 10 years                  34042.2     2946.4  11.554  < 2e-16 ***
education_levelHigh School                         -19800.6     3797.3  -5.214 1.87e-07 ***
education_levelMaster's degree                      11165.0     1097.7  10.171  < 2e-16 ***
education_levelPhD                                  36008.6     2358.9  15.265  < 2e-16 ***
education_levelProfessional degree                  60872.5     2547.9  23.891  < 2e-16 ***
education_levelSome college                        -18648.4     1931.7  -9.654  < 2e-16 ***
genderWoman                                        -23361.9     1227.8 -19.028  < 2e-16 ***
raceBlack or African American                       -9473.4     3149.2  -3.008 0.002632 **
raceHispanic                                       -17008.3     2957.1  -5.752 8.98e-09 ***
raceMiddle Eastern or Northern African             -10361.3     6305.9  -1.643 0.100382
raceNative American or Alaska Native               -17219.9     7861.6  -2.190 0.028509 *
raceWhite                                          -16440.5     1903.1  -8.639  < 2e-16 ***
regionEast                                          13015.3     1116.2  11.660  < 2e-16 ***
regionWest                                          21982.5     1265.1  17.376  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60060 on 17145 degrees of freedom
Multiple R-squared:  0.3039,    Adjusted R-squared:  0.3015
F-statistic: 129.1 on 58 and 17145 DF,  p-value: < 2.2e-16
```