Mathematical Statistics Project

Lamonarca Andrea   Menekshi Fabian

Abstract

In our project we set out to investigate how the salaries in Data Science related professions in the western region of the US are influenced by a number of variables, like education and skills, that candidates are required or advised to possess by job offers. To accomplish this task, we have carefully assembled **our own dataset** consisting of data from job offers in the data science field taken from the *indeed.com* platform. With the data collected, our objective is to train and select a simple and interpretable regression model that best captures the dependence of salary on the set of prerequisites suggested in a given job listing. Furthermore, we aim to test different hypotheses on the requirements in order to evaluate their effect on salaries and are particularly interested in analysing education levels, especially seeking an answer to the question: does a PhD guarantee a higher salary?

1 Dataset overview

To begin with, salaries in Data Science employment roles are currently undergoing significant fluctuations as the job market adjusts to the boom of AI. This increases the likelihood of income data being influenced by exogenous trends and time-dependent factors. We addressed this issue by collecting our full dataset over the span of two days to reduce the time dependence of observations and also to minimise the effects of any seasonal changes in the job demand.

We thus performed the web-scraping of *indeed.com* Data science positions using the APIFY API and then extracted the relevant variables from the job descriptions, by resorting to natural language processing (NLP) techniques. The details of these two processes are explained in Section A of the appendix and the commented code for NLP is provided separately.

1.1 The Main Variables

During the NLP of the job descriptions, we collected a total of 51 categorical variables and 2 numerical variables (SALARY and experience) with 1658 observations, which were further cleaned and combined into a single dataset. In this process, we faced a few challenges.

Firstly, in the case of education and experience variables, there were often multiple categories/values extracted for a given observation. In order to map each job listing to a single data point, we kept only the highest ranking category or numerical value (for both variables) since the fierce competition for job positions in Data Science allows only candidates that meet or exceed all the employer's expectations to succeed in obtaining the job.

Secondly, the salaries extracted via web-scraping were not uniform, meaning that some observations contained a single value, while others provided an income range and, also, many observations contained the hourly wage or monthly pay instead of the yearly salary. To address these issues, an average of the two values for the annual salary was taken and the hourly/monthly wages were converted to annual salaries by inspecting whether the job was full-time (40 hours) or part-time (24 hours, the mean of 16 and 32)[5], thus multiplying the latter pay by the number of working hours in a year. We speculated this straightforward approach was probably not ideal since it involved extrapolating yearly salaries from monthly or even hourly payments, and, as working hours were not always specified, this method created nearly 100 missing values.

The list of variables in the final dataset grouped by their general class is:

id: a variable matching the last part of the URL of each job offer identifying it uniquely.

Location: a categorical variable indicating the US state for a given job.

Experience: a numerical variable containing the highest number of years of professional experience suggested by employers in the job description.

HighestEducation: a categorical variable showing the highest level of education suggested by employers in the job description.
ProgrammingSkills: binary categorical variables indicating the presence of a given programming skill in the job offer.
DataScienceSkills: binary categorical variables indicating the presence of a given technical skill in the job offer.
SoftSkills: binary categorical variables indicating the presence of soft skills in the job offer.
Jobtype: categorical variable indicating the type of work performed.
Seniority: categorical variable indicating the seniority level.
SALARY: continuous variable indicating the average yearly salary offered by each job.

2 Exploratory Data Analysis

First, we refined our previous data cleaning by discarding rows whose SALARY was missing and removing two categorical skill variables, `dart` and `stress management`, since they were found to be zero for all observations. Then, it was noticed that the `HighestEducation` variable contained a category, "Associate's degree", with no matching observations and so it was dropped. Analysing the `id` variable, the number of unique entries matched the dataset's size, proving that no duplicates were present and so `id` was removed since it was no longer needed.
Then we started to analyse graphically the relationships between the covariates and the response, SALARY, to evaluate their effects. By examining the boxplots and violinplots (with Gaussian KDE) of `Jobtype`, `Location` and `HighestEducation` against salary (figures 2,3,4 in Appendix), we found that the variables' different categories seem to affect quite considerably the corresponding salary distributions. The next step was to inspect the distribution of the response, SALARY, due to its importance for the linear regressions we intended to use.

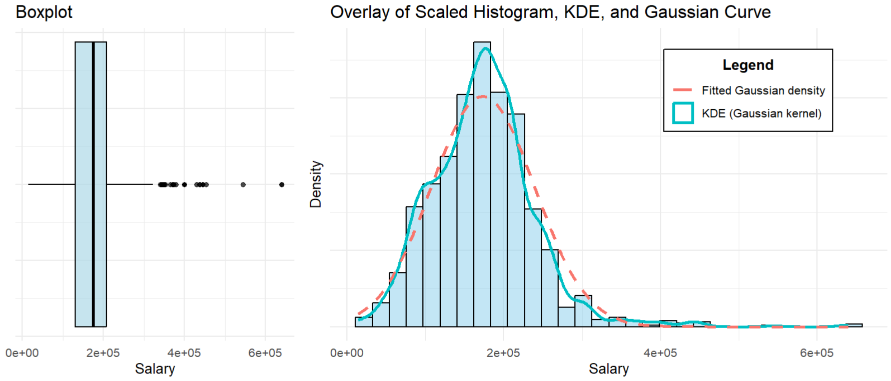


Figure 1: Boxplot and Scaled histogram vs KDE vs Fitted Gaussian density for SALARY

In Figure 1, it can be concluded from both the boxplot and the histogram that the distribution is mostly symmetric and fits modestly well the Gaussian plot despite a few significant outliers in the right tail. The observations associated with these extreme salaries were further investigated and no collected variable seemed to justify the results as almost all skills were absent and there was no predominant jobtype, education or location. Tracing back these job offers to the original dataset, it was concluded that these salaries corresponded mainly to job descriptions with hourly wages, which were extrapolated during NLP, showing that the above extrapolation procedure clearly failed to be optimal. Since these observations very likely provided incorrect/misleading information that would violate the normality assumption in the regression models, as seen from QQ-plots (figure 8 in Appendix), the associated 29 rows were removed.

3 Naive Analysis

3.1 Multivariate linear Regression

To begin with, we used all variables to predict SALARY, fitting a multiple linear regression on the whole dataset and evaluated the results. With this simple approach, we obtained a linear model for which $R^2 = 0.439$, adjusted $R^2 = 0.414$ and which, according to the RMSE (Root mean square error), was on average off by approximately 42,500 \$ in predicting SALARY. The residuals of the regression were approximately normally distributed with zero mean as observed from histogram in figure 12 in Appendix. The normality assumption was further tested by running a Shapiro-Wilk test on the residuals and the latter rejected the normality assumption at significance level 0.05. However, given the large sample size, the test can detect even small deviations from normality and, since the distribution was approximately symmetric with (seemingly) exponentially decaying tails, the test result should not be considered a strong violation. Moreover, there didn't seem to exist sufficient evidence to support any strong violation of the homoscedasticity assumption, as was noticed by plotting residuals against fitted regression values and against covariates like `Experience` (figures 12, 14 in Appendix).

Pursuing to better understand the dependence of SALARY on the covariates, we analysed the results of the t-tests of the form $H_0 : \beta_k = 0$ for the regression coefficients $\beta_k : 1 \leq k \leq p$ of all the variables (in total 64 - including dummy variables for levels of categorical variables). The variables for which H_0 was rejected with $\mathrm{p-value} < 0.05$ were only 28 indicating that the majority of the requirements were not relevant for the model's predictions. The features found to be significant most notably comprised `Experience`, all levels of `Location`, some levels of `HighestEducation` and `Jobtype`, programming skills like `sql`, `c` and `javascript` as well as soft skills like `flexibility`, `leadership` and `attention to detail`.

Accordingly, we expected a small number of predictors to be much more important in determining SALARY and pursued to test our expectations, by selecting the most relevant features, using LASSO regression and STEP-DOWN/UP methods. In this way, we also try to reach our primary goal of building a simple regression model with which to interpret SALARY dependencies.

3.2 Box-Cox transformation

In our analysis, we have and will focus on linear models, but, in this subsection, we try to check the underlying assumption that the "true"/"right" model for the data is linear. Thus, we transformed the SALARY variable in order to evaluate if a polynomial/exponential regression was better suited to describe salary's growth based on the collected requirements. We experimented with multiple transformations in the Box-Cox family, for values of λ between 0 and 5 in increments of 0.1, fitting every time a linear model with all covariates. To evaluate the linear fits we relied upon R^2 since the latter is independent of the scale of the predicted variable. We found that the maximum value of R^2 was 0.447 for $\lambda = 0.5$. As this consisted in a mere 1% rise in R^2, we disregarded the SALARY transformation because there didn't seem to exist a sufficient increase in performance to violate our assumption.

4 Feature selection

4.1 LASSO and ELASTIC NET regression

For the reasons explained above, we now trained a LASSO regression using 10-fold cross-validation (on the whole dataset) so as to find the optimal λ for the L^1 penalty. The latter model sets with high probability the coefficients of irrelevant variables to zero thanks to its L^1 penalty and so can be used to simplify the relation between design matrix and SALARY. LASSO (with optimal λ) didn't offer significant improvements in predictive performance compared to the naive model showcasing RMSE $\approx 43,000\$$, $R^2 = 0.435$ and adjusted $R^2 = 0.415$. Remarkably, to achieve the same performance, LASSO used 11 less predictors and the selected features mostly aligned with the results of the t-tests in section 3.1: `Jobtype`, `Experience`, `Location`, `HighestEducation` along with previously not significant variables like `Seniority`.

In an attempt to improve the predictions of LASSO, we implemented the wider class of ELASTIC NET regressions, which uses a weighted average of L^1 and L^2 penalties. In R, for the vector of parameters β, the latter model is implemented with only one parameter $0 \leq \alpha \leq 1$ so that:

$$\text{ELASTIC NET penalty} = \alpha \cdot \|\beta\|_1 + (1-\alpha) \cdot \|\beta\|_2^2$$

Since there was no reason to suppose that the model associated to any α should perform better than another, we fitted all models from $\alpha = 0$ (Ridge regression) to $\alpha = 0.95$ (almost LASSO) in increments of 0.05. For each model, we computed the usual metrics and observed that the latter were maximised (or minimised, for RMSE) for different values of α: according to both the R^2 and RMSE, the preferred model was the Ridge regression whereas the value of α which maximised the adjusted R^2 was the LASSO regression. Examining more closely the results $\forall \alpha$ together with the LASSO results, it appears evident that the margin by which RIDGE was preferred for the RMSE and R^2 compared to LASSO was negligible. This supports the overall superiority of LASSO among ELASTIC NET models, offering the best compromise between interpretability and good predictions.

4.2 Step-up and step-down models with BIC

To further evaluate which variables are significant for SALARY, we implemented step-down and step-up linear models based on the Bayesian information criterion (BIC) using built-in R functions. It was decided to rely on BIC instead of the Akaike information criterion in light of our goal of finding the "true" and possibly simplest model for SALARY. The BIC method adds a penalty based on model size to the maximised likelihood function to prevent overfitting and, in R, for a model with k parameters and n observations, the latter is given by $\frac{1}{2}k \log n$. The step-up and step-down linear models both achieved adjusted $R^2 = 0.4$ and agreed in selecting the same covariates, chiefly `Location`, `EXPERIENCE`, `Jobtype`, `HighestEducation`, soft skills like `leadership` and `attention to detail` and hard skills such as `Computer vision` as well as `javascript`, validating the findings of the above sections.

Moreover, due to the categorical nature of most variables, we believed interactions between them to be quite relevant. We tested our conjecture by adding a sensible interaction term to the linear regression with all covariates, namely `EXPERIENCE : Seniority` (seniority weighted by Experience) as jobs with higher seniority require more experience. This model showed substantial improvements in R^2 and RMSE, as expected from the increase in the number of covariates, but also in the adjusted R^2 by nearly 2%. In addition, one of the p-values for the t-tests on the regression coefficients of the 2 newly included variables was significant at a level close to zero. Hence, we implemented a step-up method to select the best model with binary interactions between the covariates based on BIC. Starting from the linear model in section 3, we iterated through all possible interaction terms fitting every time a linear model with all covariates and the previously accepted interactions. We then checked whether the BIC of this new linear model decreased and if so, we included the interaction term in the current best model. The selected linear model included a total of 12 interaction terms out of the $\binom{49}{2} = 1176$ possible, some of which were: `Seniority:Experience`, `Computer vision:python` and `Deep learning:java` showcasing the synergies between ML knowledge and programming skills as well as between responsibility and experience. This linear model improved the adjusted R^2 up to 0.464.

5 Hypothesis testing

In this section, we will analyse which job requirements mainly affect SALARY, focusing on the impacts of technical skills, soft skills and the education level. Firstly, asymptotic t-tests were performed individually for all skills, testing the one-sided alternative hypothesis $H_1 : \mu_1 > \mu_0$: μ_1 = mean salary of class 1 (skill is required), μ_0 = mean salary of class 0. These showed that 13 out of the 42 tested variables had a statistically significant and positive impact on salary at level 0.05. The significant features included mainly technical skills such as `Computer vision`, `NLP`, `Deep Learning`, `Scala` and `Java`, `C++`, underlining the higher retribution for advanced pro-

gramming and ML skills. The soft skills found to be relevant included `creativity`, `leadership`, `collaboration` and `conflict resolution`. These findings point out that both technical proficiency and interpersonal skills are important in reaching higher salaries.
Furthermore, education levels were found to significantly influence SALARY. In fact, the F-test from ANOVA for $H_0 : \alpha = 0$, where α denotes the main effect of `HighestEducation`, returned a p-value ≈ 0, rejecting H_0 and thus confirming its relevance in affecting SALARY. These results were then strengthened by the Tukey's honest significance test, which showed that "Doctorate" degree holders earned a mean salary much different than those with "Bachelor's Degree" and "Master's Degree", both with $p < 0.001$, even if there was no significant difference between Bachelors and Masters ($p = 0.65$). We then focused on the PhD category testing it against all other classes, using an asymptotic t-test such that $H_1 : \mu_1 > \mu_0 : \mu_1 =$ mean salary of PhD, $\mu_0 =$ mean salary of all other classes. Its positive impact on the average salary was reported to be statistically significant with $p \approx 0$. Then, the same t-test for the mean salary was repeated for PhD against the means of all other classes separately. These tests showed that PhDs brought a statistically significant (at level 0.001) increase in salary when compared to both Bachelor's or Master's Degree, but there was insufficient evidence to draw a strong conclusion for jobs with no formal education required ($p \approx 0.4$). Based on job offers included in this category, we speculate that the result reflected positions where skills were more important than formal qualifications.

6 Conclusion and Discussion

Overall, using different linear regression models, we found and confirmed that many of the collected job requirements are unnecessary to predict salary. Thus, a small group of covariates, chiefly `Experience`, `HighestEducation`, `Location`, and `Jobtype` along with a few soft and technical skills, dominates in affecting earnings, and the combined impact on income of a few variables like `Seniority:Experience` is not to be neglected either. Through F-tests and t-tests, we have also concluded that Education affects salary and Doctorates are statistically significant for gaining higher salaries. However, these methods are inherently limited to establishing correlations between variables and cannot, on their own, demonstrate causality. On top of that, our analysis still has some other important limitations:

- The dataset likely contained some errors inherent to the NLP techniques used for data extraction, but, given this was our first attempt at NLP, we regard the dataset to be satisfactory.
- Our findings for metrics like R^2 are ambiguous in light of overfitting problems, as the same dataset was used for both model training and evaluation. However, this decision to forego data splitting was taken to prioritise interpretability and owing to the dataset's nature. In fact, given the modest size of the dataset compared to the number of predictors, when we tried splitting the dataset for training (80%) and testing (20%), we soon realised that variables with few observations were completely missing from either dataset. Moreover, the size of the testing set was comparable to the total number of variables (including dummies) causing the adjusted R^2 to be overly influenced by parameters' count and not reflecting the proportion of explained variance. Furthermore, the larger sample size reduced the chances of a type 2 error in hypothesis testing for both linear models and ANOVA in sections 3.1 and 5. We thus managed to draw more conclusions about the covariates, given the same type 1 error threshold.
- The overall size of tests on Education in Section 5 is unclear, but, given the small p-values, even upon applying a Bonferroni correction, they still remain significant at level 0.05. Also, the normality assumption on the different splits of SALARY, which underlies all the tests used, couldn't be adequately checked. However, given the large sample size, it was assumed that the Central limit theorem would apply and, by examining the violin-plots in figure 2, the KDEs don't qualitatively differ substantially from a Gaussian density.

It is still possible to improve our understanding of how salary depends on requirements by:

- Collecting a dataset with more observations so as to perform meaningful cross-validation to verify the evaluation metrics reported above as well as to balance the observations' count in the categories of different variables like Location, thus reducing any bias towards a given class.

- Training a non-linear or non-parametric regression to capture more complex trends that are not well-represented by linear models, thus checking our conclusion to section 3.2.
- Converting the regression problem into a classification one by discretizing the salary variable into a given number of quantiles. Then, one could use a logistic regression to predict the right salary range for a given set of job requirements and this would probably simplify the prediction problem leading to higher performance.

References

[1] Fetsje Bijma, Marianne Jonker, and Aad van der Vaart. *An introduction to Mathematical Statistics*. ISBN: [978-94-6298-510-0]

[2] Apify Technologies s.r.o.. *Apify API*. URL: https://apify.com

[3] Ondra Urban. *APIFY API: "Is web scraping legal?"*. URL: https://blog.apify.com/is-web-scraping-legal/

[4] Indeed.com. *Data Science Jobs, Employment*. URL: https://www.indeed.com/jobs?q=data+science

[5] U.S. Department of Commerce. *Office of Human Resources Management: "Hours of duty and work schedules"*. URL: https://www.commerce.gov/hr/practitioners/leave-policies/hours-of-duty-and-work-schedules#:~:text=Part%2Dtime%20Employment.&text=A%20permanent%20part%2Dtime%20employee,not%20held%20to%20these%20limitations

[6] STHDA (Statistical tools for high-throughput data analysis). *ggplot2 - Essentials*. URL: https://www.sthda.com/english/wiki/ggplot2-essentials

[7] Datacamp. *RDocumentation*. URL: https://www.rdocumentation.org

[8] spaCy. *Industrial-Strength Natural Language Processing - In Python*. URL: https://spacy.io

[9] Data with Rez (YouTube). *Easy Data Science Project: Extracting information from job description*. URL: https://youtu.be/3LAY7rocJmg?si=nPp-Rl6b71uEB81j

[10] Dr. William Mattingly - freeCodeCamp.org (YouTube). *Natural Language Processing with spaCy & Python - Course for Beginners*. URL: https://youtu.be/dIUTsFT2MeQ?si=XbiPUWiIqxv4XatG

Appendix

A Dataset's creation and content

A.1 Web Scraping

We navigated the *indeed.com*[4] platform and performed a search for US states on the Western coast labeled under the Data Science field. We then resorted to a pre-existing API in the APIFY[2] website to scrape the relevant Indeed website pages and collected all data available for the states listed in section 1.3, exporting it as an excel file. The web scraping performed by the APIFY API gathers only publicly accessible data, while avoiding personal information and is thus compliant with ethical and legal regulations, such as the CCPA and other federal laws enforced in the US.[3]

A.2 NLP for Features' Extraction

We then began to analyse the excel file removing duplicate rows and rows with missing salary. Since most of the data relevant to our analysis was stored inside the job description text or in other columns as strings, we extracted most of the features in the final dataset by employing different natural language processing techniques[1]. We employed 3 main types of functions/heuristics:

- A key-word search through the text of the job description using a dictionary to match the keywords to the corresponding category extracting for each row a label (for features like education and programming skills);
- Matching groups of tokens in the text to key-words, values in a dictionary and then to the corresponding categories. This was acheived by stemming words in the text and comparing them to the stemmed words in the dictionary or by selecting the category with the highest cosine similarity (a metric commonly used to evalaute how similar two words are) by utilizing the word embeddings of Word2Vec created by Google (for features like seniority and jobtype);
- Training a custom Spacy model using a small labelled dataset with 500 observations (taken from [9]) to customise the Named Entity Recoignition labels of and thus extract info like experience, skills, ... from the job description (used for the experience feature mainly).

A.3 Variables

All variable names (apart from id) and levels of categorical variables are listed below:
HighestEducation = [None, Associate's Degree, Bachelor's Degree, Master's Degree, Doctorate]
ProgrammingSkills = [python, r, sql, julia, scala, java, c++, matlab, sas, javascript, ruby, php, perl, Swift, kotlin, shell, dart, c, rust]
DataScienceSkills = [NLP_software, Deep_learning, Compuer_vision, database_software, visualisation_software, general_DS_software]
SoftSkills = [communication, teamwork, problem solving, creativity, adaptability, time management, critical thinking, leadership, collaboration, interpersonal skills, conflict resolution, emotional intelligence, organization, negotiation, decision making, active listening, flexibility, attention to detail, self-motivation, motivation, stress management]
Jobtype = [Data analyst, Software Engineer, Researcher, ML/AI Engineer, Manager, Statistician, Sales, Cloud ops, Other]
Seniority = [Not specified, Junior, Senior]
Location = [California, Washington, Oregon, Idaho, Nevada, Arizona, Utah]

[1] for convenience of the available libraries the NLP was implemented in python and the code has been uploaded and commented to prove our first-hand work

B Plots from EDA

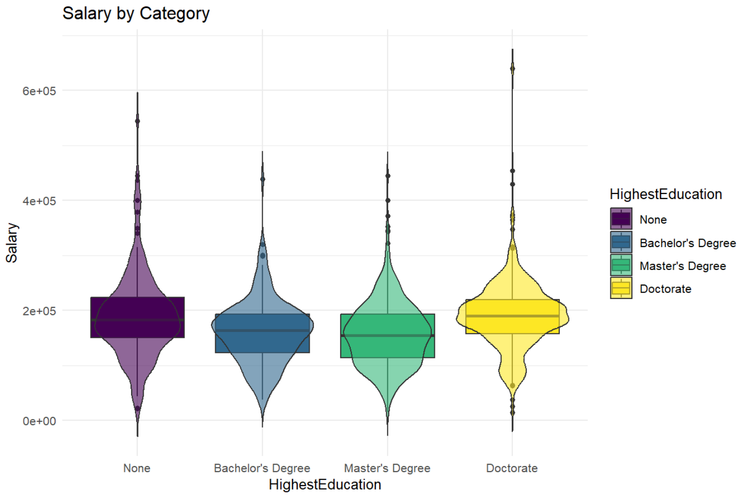


Figure 2: Boxplot and Violiplot of salary sorted by education categories

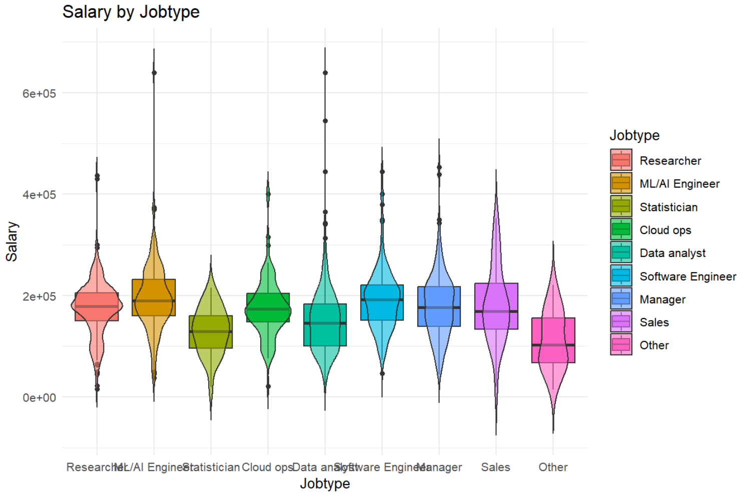


Figure 3: Boxplot and Violiplot of salary sorted by jobtype categories

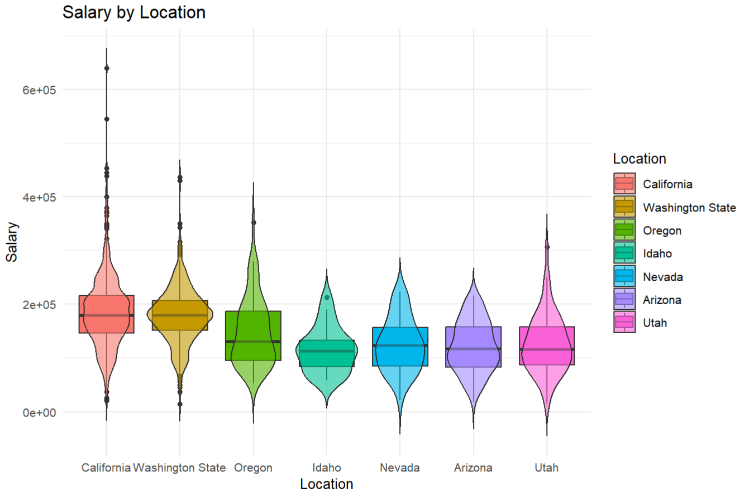


Figure 4: Boxplot and Violiplot of salary sorted by location categories

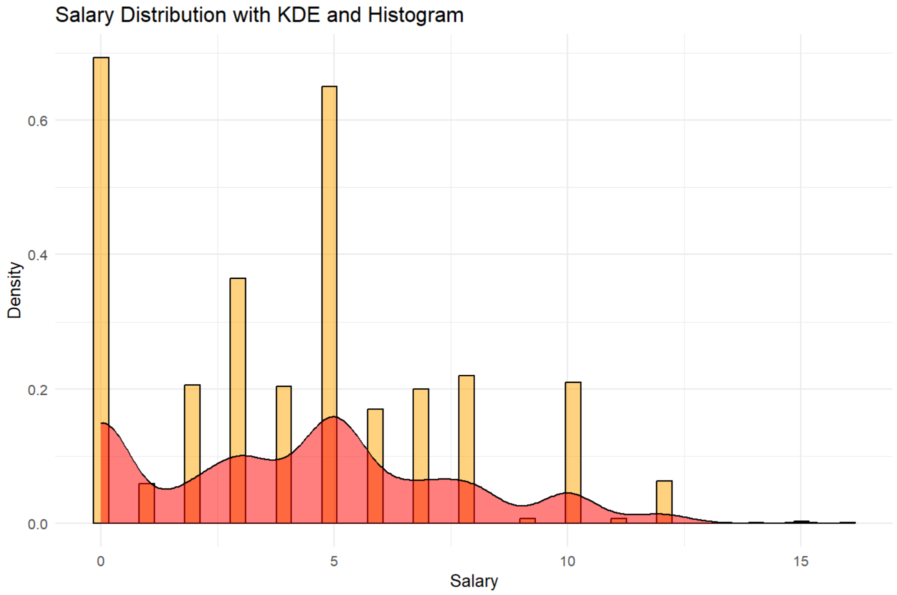


Figure 5: KDE with Gaussian kernel against histogram of years of Experience

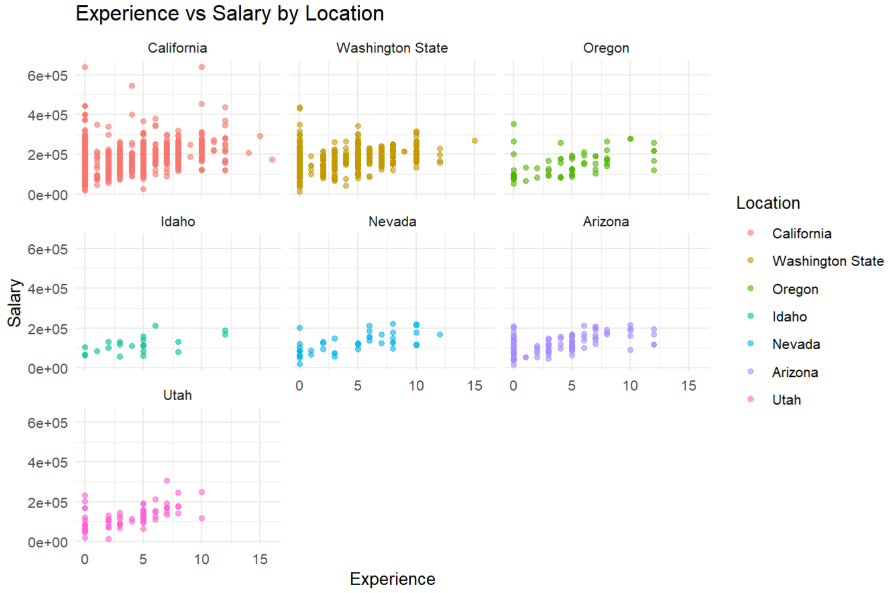


Figure 6: Scatter plot of salary vs experience sorted by location

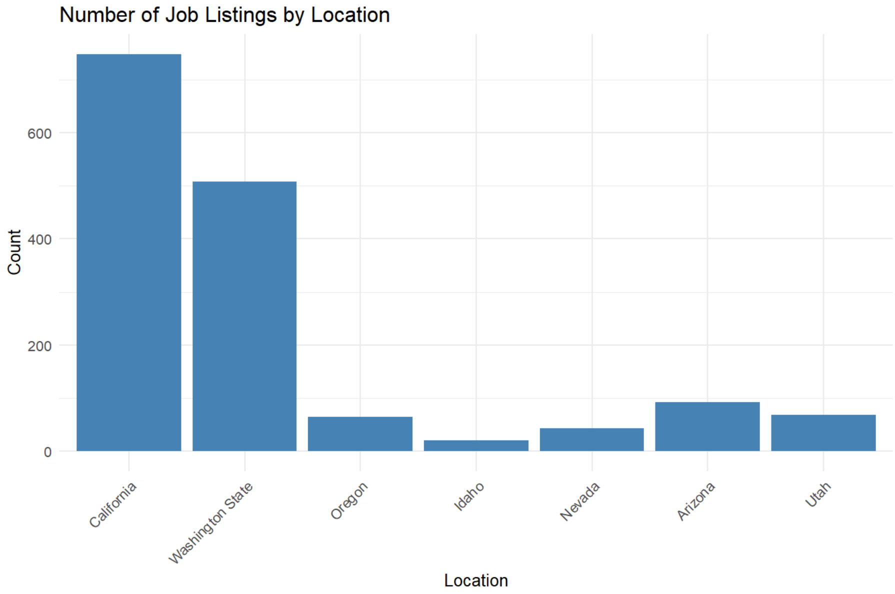


Figure 7: Scatter plot of salary vs experience sorted by location

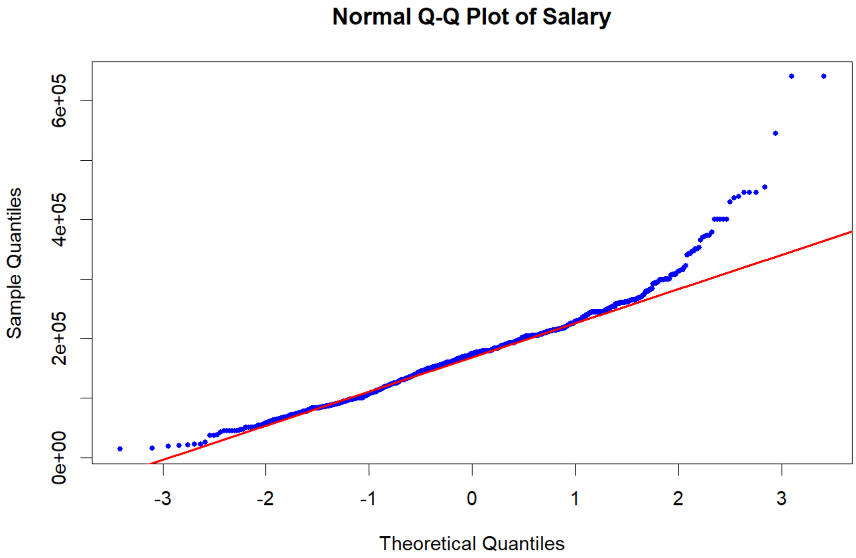


Figure 8: QQ-plot with normal quantiles of Salary for dataset with outliers

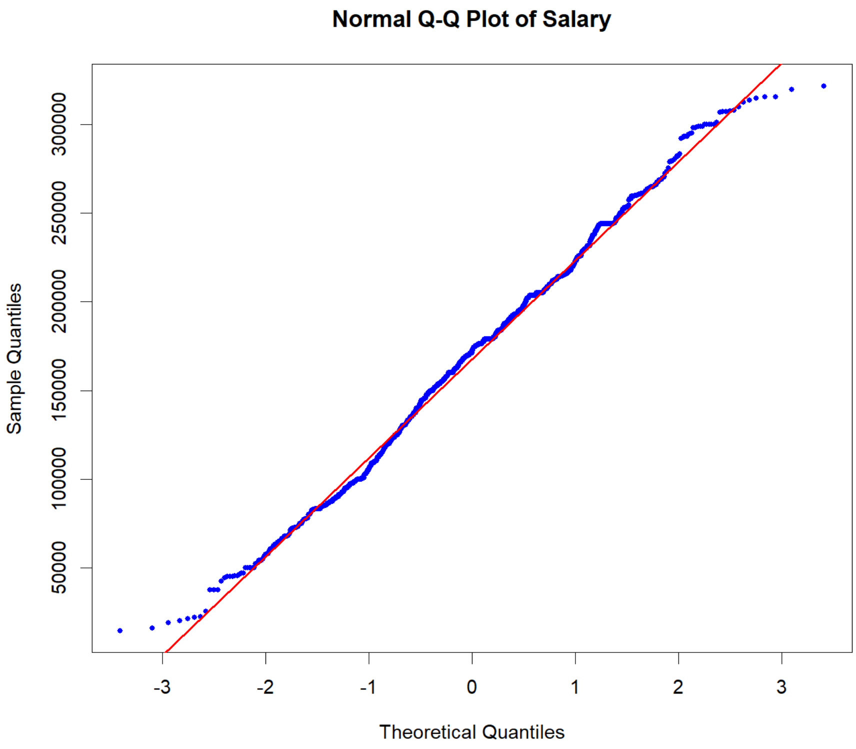
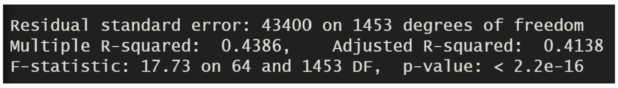


Figure 9: QQ-plot with normal quantiles of Salary without outliers

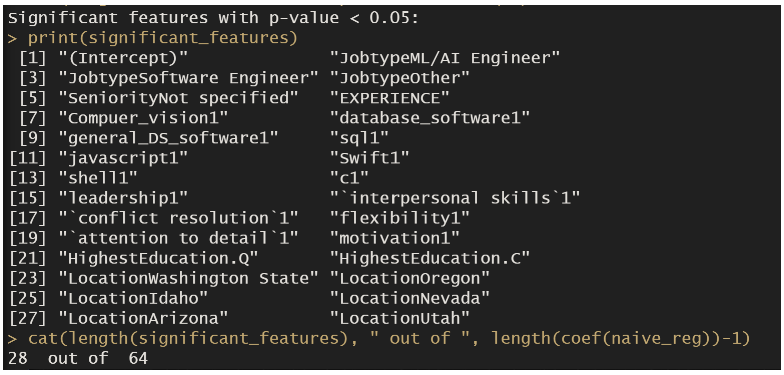
C Plots and output of Naive analysis

The output of the t-tests for all the variables is missing because it could not be captured adequately by the screenshot, but our results can be confirmed by running the R code.



```
Residual standard error: 43400 on 1453 degrees of freedom
Multiple R-squared:  0.4386,	Adjusted R-squared:  0.4138 
F-statistic: 17.73 on 64 and 1453 DF,  p-value: < 2.2e-16
```

Figure 10: Output of multiple linear regression



```
Significant features with p-value < 0.05:
> print(significant_features)
 [1] "(Intercept)"                "JobtypeML/AI Engineer"     
 [3] "JobtypeSoftware Engineer"  "JobtypeOther"              
 [5] "SeniorityNot specified"    "EXPERIENCE"                
 [7] "Compuer_vision1"           "database_software1"        
 [9] "general_DS_software1"      "sql1"                      
[11] "javascript1"               "Swift1"                    
[13] "shell1"                    "c1"                        
[15] "leadership1"               "`interpersonal skills`1"   
[17] "`conflict resolution`1"    "flexibility1"              
[19] "`attention to detail`1"    "motivation1"               
[21] "HighestEducation.Q"        "HighestEducation.C"        
[23] "LocationWashington State"  "LocationOregon"            
[25] "LocationIdaho"             "LocationNevada"            
[27] "LocationArizona"           "LocationUtah"              
> cat(length(significant_features), " out of ", length(coef(naive_reg))-1)
28  out of  64
```

Figure 11: Variables whose parameters are significant in linear regression

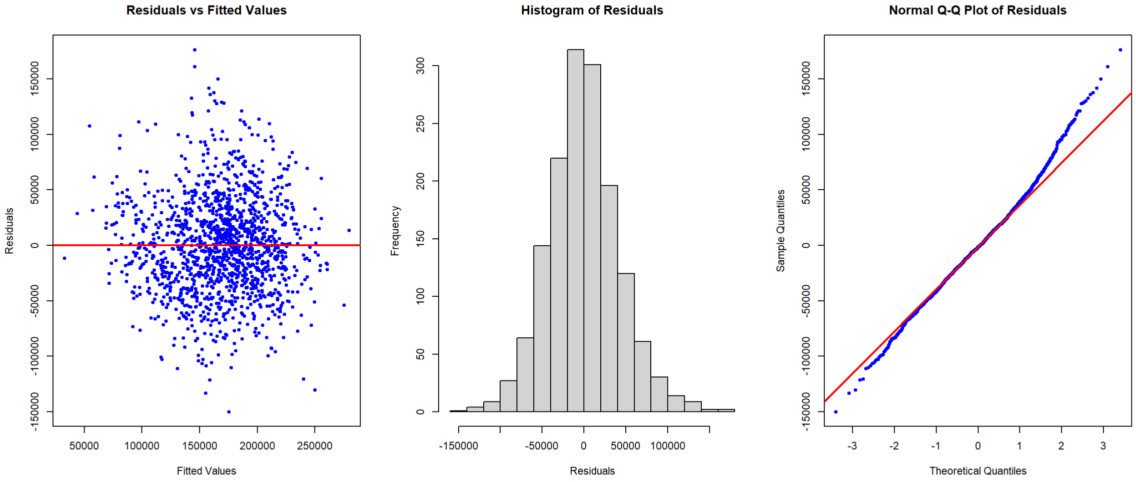


Figure 12: Diagnostics of residuals of multivariate regression

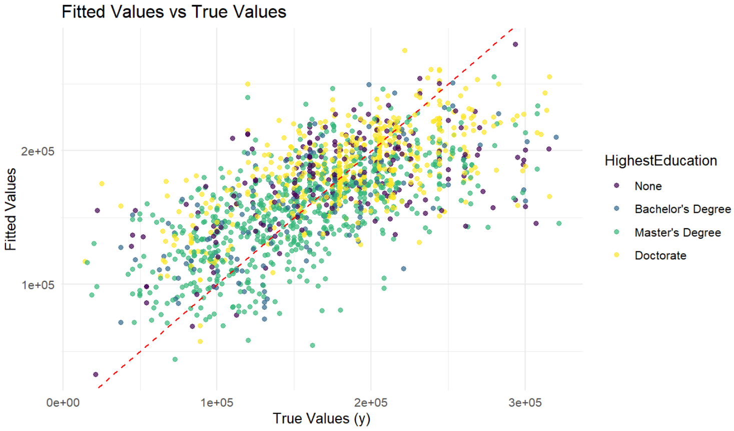


Figure 13: Fitted regression values vs true values for SALARY colour-coded by HighestEducation

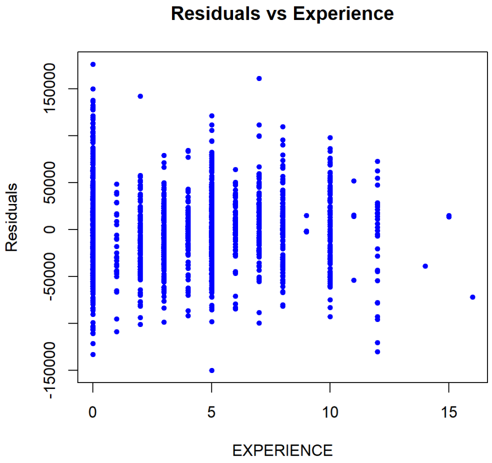


Figure 14: Scatter plot of residuals vs Experience for homoscedasticity assumption

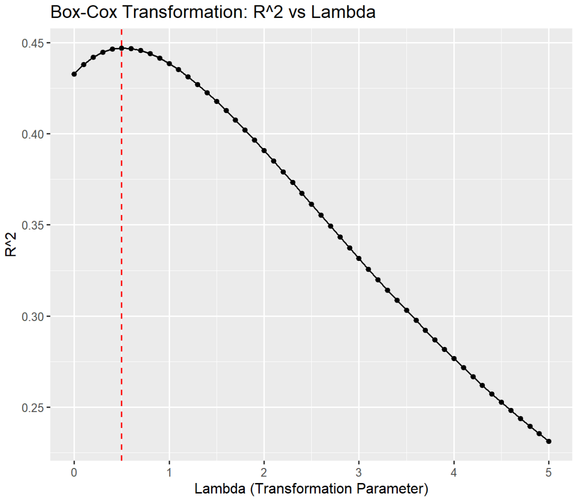


Figure 15: Box-Cox transformation

D Plots and Output of Feature selection

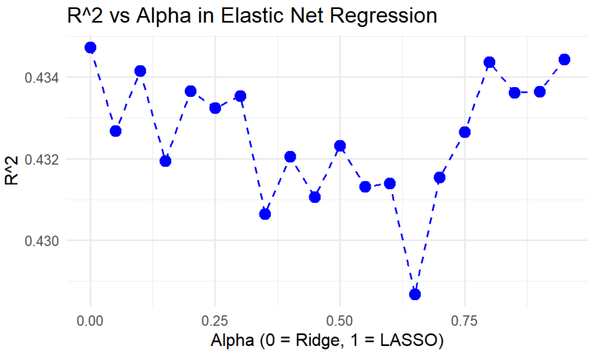


Figure 16: α vs R^2 for ELASTIC-NET

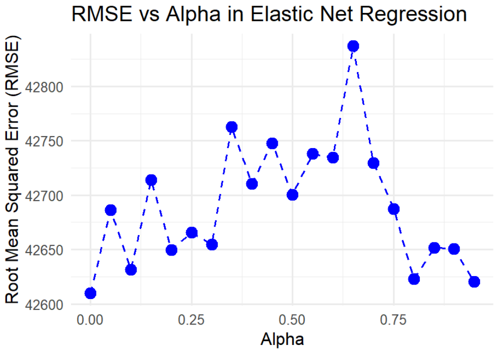


Figure 17: α vs RMSE for ELASTIC-NET

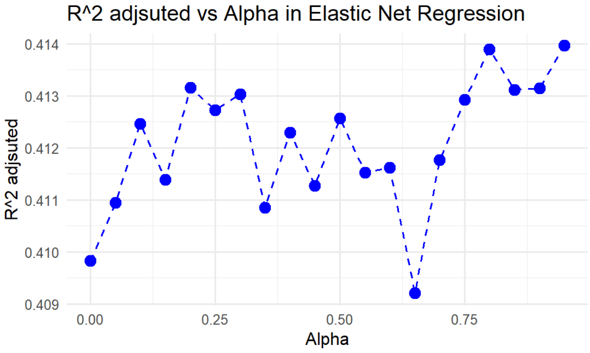
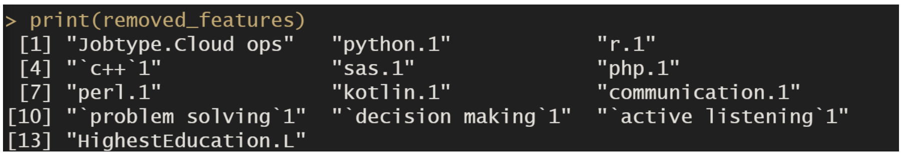
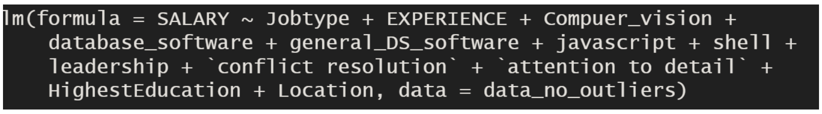


Figure 18: α vs adjusted R^2 for ELASTIC-NET



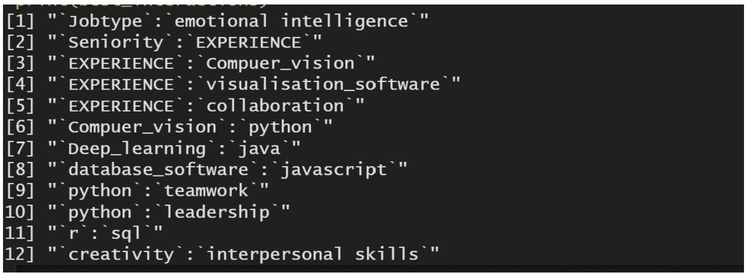
```
> print(removed_features)
 [1] "Jobtype.Cloud ops"    "python.1"            "r.1"
 [4] "`c++`1"               "sas.1"               "php.1"
 [7] "perl.1"               "kotlin.1"            "communication.1"
[10] "`problem solving`1"   "`decision making`1"  "`active listening`1"
[13] "HighestEducation.L"
```

Figure 19: Variables whose parameters are set to zero by LASSO regression



```
lm(formula = SALARY ~ Jobtype + EXPERIENCE + Compuer_vision +
    database_software + general_DS_software + javascript + shell +
    leadership + `conflict resolution` + `attention to detail` +
    HighestEducation + Location, data = data_no_outliers)
```

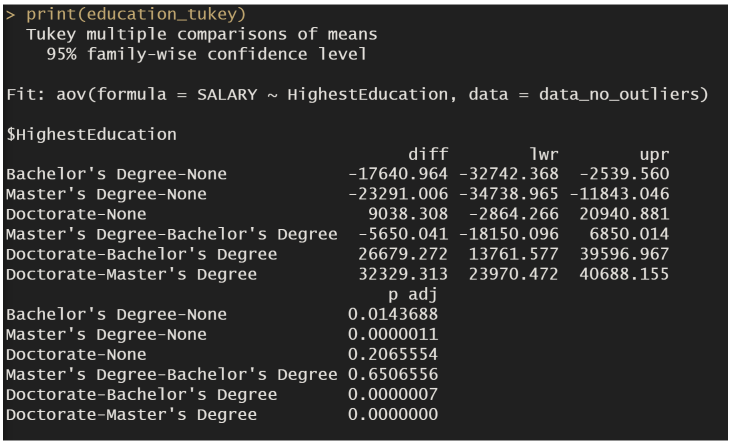
Figure 20: Variables selected by BIC step-up and step-down models



```
[1] "`Jobtype`:`emotional intelligence`"
[2] "`Seniority`:`EXPERIENCE`"
[3] "`EXPERIENCE`:`Compuer_vision`"
[4] "`EXPERIENCE`:`visualisation_software`"
[5] "`EXPERIENCE`:`collaboration`"
[6] "`Compuer_vision`:`python`"
[7] "`Deep_learning`:`java`"
[8] "`database_software`:`javascript`"
[9] "`python`:`teamwork`"
10] "`python`:`leadership`"
11] "`r`:`sql`"
12] "`creativity`:`interpersonal skills`"
```

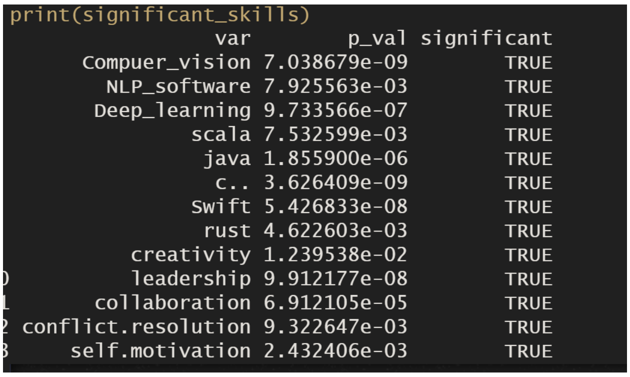
Figure 21: Interaction terms selected by BIC step-up model

E Hypothesis testing output



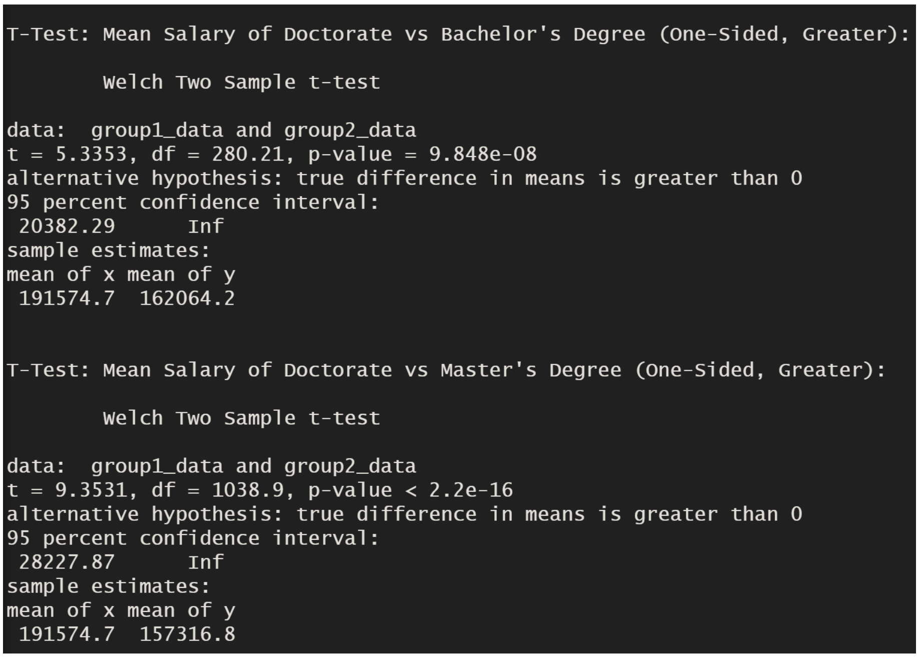
```
> print(education_tukey)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SALARY ~ HighestEducation, data = data_no_outliers)

$HighestEducation
                                        diff        lwr        upr
Bachelor's Degree-None            -17640.964 -32742.368  -2539.560
Master's Degree-None              -23291.006 -34738.965 -11843.046
Doctorate-None                      9038.308  -2864.266  20940.881
Master's Degree-Bachelor's Degree  -5650.041 -18150.096   6850.014
Doctorate-Bachelor's Degree        26679.272  13761.577  39596.967
Doctorate-Master's Degree          32329.313  23970.472  40688.155
                                      p adj
Bachelor's Degree-None            0.0143688
Master's Degree-None              0.0000011
Doctorate-None                    0.2065554
Master's Degree-Bachelor's Degree 0.6506556
Doctorate-Bachelor's Degree       0.0000007
Doctorate-Master's Degree         0.0000000
```

Figure 22: Tukey's honest significance test on HighestEducation



```
print(significant_skills)
                  var        p_val significant
       Compuer_vision 7.038679e-09        TRUE
         NLP_software 7.925563e-03        TRUE
        Deep_learning 9.733566e-07        TRUE
                scala 7.532599e-03        TRUE
                 java 1.855900e-06        TRUE
                  c.. 3.626409e-09        TRUE
                Swift 5.426833e-08        TRUE
                 rust 4.622603e-03        TRUE
           creativity 1.239538e-02        TRUE
0          leadership 9.912177e-08        TRUE
1       collaboration 6.912105e-05        TRUE
2 conflict.resolution 9.322647e-03        TRUE
3     self.motivation 2.432406e-03        TRUE
```

Figure 23: Significant skills for one-sided asymptotic t-tests



```
T-Test: Mean Salary of Doctorate vs Bachelor's Degree (One-Sided, Greater):

        Welch Two Sample t-test

data:  group1_data and group2_data
t = 5.3353, df = 280.21, p-value = 9.848e-08
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 20382.29      Inf
sample estimates:
mean of x mean of y
 191574.7  162064.2


T-Test: Mean Salary of Doctorate vs Master's Degree (One-Sided, Greater):

        Welch Two Sample t-test

data:  group1_data and group2_data
t = 9.3531, df = 1038.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 28227.87      Inf
sample estimates:
mean of x mean of y
 191574.7  157316.8
```

Figure 24: One-sided asymptotic t-tests for Doctorate against Bachelor's and Master's degree