Internship Project Title	RIO-125: HR Salary Dashboard Train dataset and prediction
Name of the Company	TCS iON
Name of the Industry Mentor	Debashis Roy
Name of the Institute	ICT ACADEMY OF KERALA

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
1/05/2023		10.5	Chrome, Windows 10	MS Office, Python

Acknowledgements

I am highly indebted to TCS iON for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.
I would like to express my gratitude towards my parents and my academic mentor, for their kind co-operation and encouragement which help me in completion of this project. I would like to express my special gratitude and thanks to my industry mentor for giving me such attention and time.

Objective

This HR data is a fictional data set created by Dr. Carla Patalano and Richard Hubenar. HR data can be hard to come by, and HR professionals generally lag behind with respect to analytics and data visualization competency. The data set is to teach HR students how to use and analyze the data in Tableau Desktop - a data visualization tool that's easy to learn.

In this analysis process, I will use python models such as logistic regression, Decision tree to train and predict thee model with highest accuracy will be used for prediction. I will also create a dash board using PowerBI.

Introduction

The Dataset revolves around a fictitious company and the core data set contains names, DOBs, age, gender, marital status, date of hire, reasons for termination, department, whether they are active or terminated, position title, pay rate, manager name, and performance score.
In this project, we will create a Hr dashboard and prediction of salaries based on Gender, Department, Employee ID and Performance Score

Overview

Employee Name : Employee’s full name

EmpID : Employee ID is unique to each employee
MarriedID : Is the person married (1 or 0 for yes or no)
MaritalStatusID : Marital status code that matches the text field MaritalDesc
EmpStatusID : Employment status code that matches text field EmploymentStatus
DeptID : Department ID code that matches the department the employee works in
PerfScoreID : Performance Score code that matches the employee's most recent performance score
FromDiversityJobFairID : Was the employee sourced from the Diversity job fair? 1 or 0 for yes or no
PayRate : The person's hourly pay rate. All salaries are converted to hourly pay rate
Termd : Has this employee been terminated - 1 or 0
PositionID : An integer indicating the person's position
Position : The text name/title of the position the person has
State : The state that the person lives in
Zip : The zip code for the employee
DOB : Date of Birth for the employee
Sex : Sex - M or F
MaritalDesc : The marital status of the person (divorced, single, widowed, separated, etc)
CitizenDesc : Label for whether the person is a Citizen or Eligible NonCitizen
HispanicLatino : Yes or No field for whether the employee is Hispanic/Latino
RaceDesc : Description/text of the race the person identifies with
DateofHire : Date the person was hired
DateofTermination : Date the person was terminated, only populated if, in fact, Termd = 1
TermReason : A text reason / description for why the person was terminated
EmploymentStatus : A description/category of the person's employment status. Anyone currently working full time = Active
Department : Name of the department that the person works in
ManagerName : The name of the person's immediate manager
ManagerID : A unique identifier for each manager.
RecruitmentSource : The name of the recruitment source where the employee was recruited from
PerformanceScore : Performance Score text/category (Fully Meets, Partially Meets, PIP, Exceeds)
EngagementSurvey : Results from the last engagement survey, managed by our external partner
EmpSatisfaction : A basic satisfaction score between 1 and 5, as reported on a recent employee satisfaction survey
SpecialProjectsCount : The number of special projects that the employee worked on during the last 6 months
LastPerformanceReviewDate : The most recent date of the person's last performance review.
DaysLateLast30 : The number of times that the employee was late to work during the last 30 days

Summary:

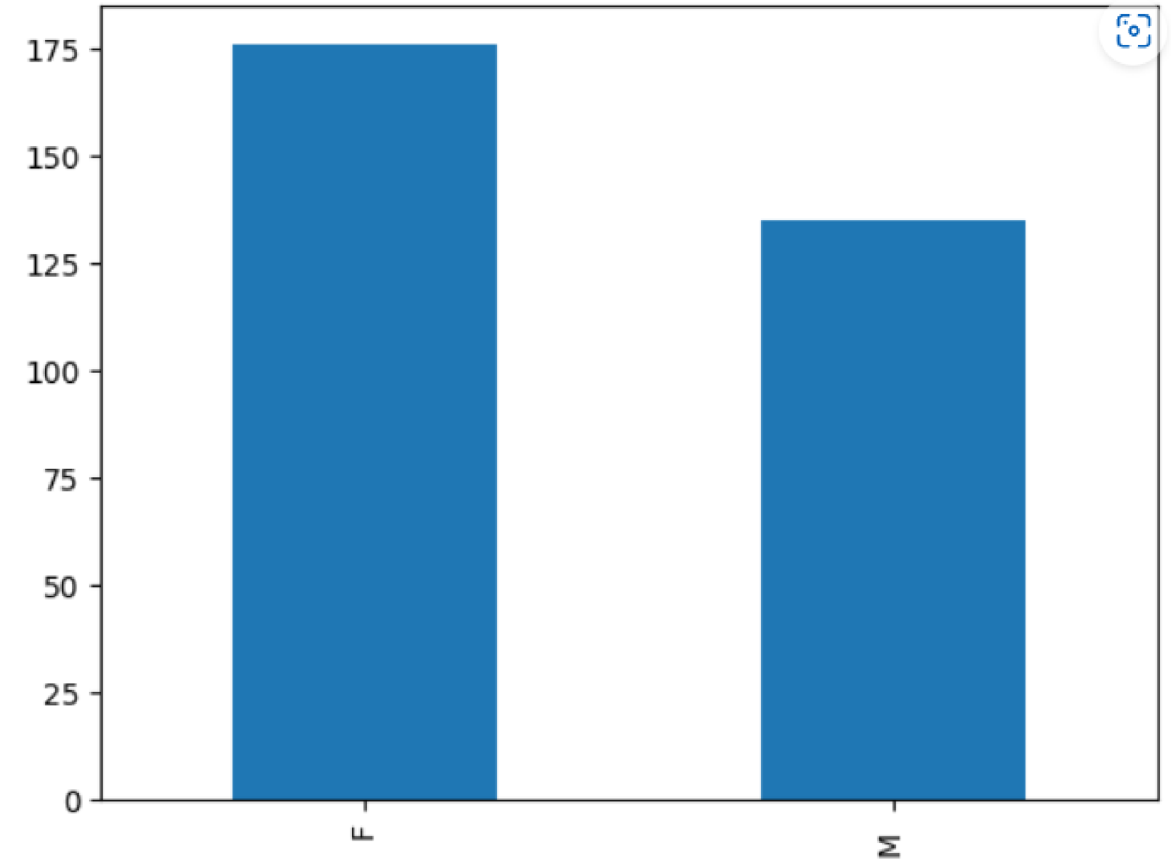
Dataset Structure: RangeIndex: 311 entries, 0 to 310
Missing Data: DateofTermination has 207 missing values
Data Type: We only have two datatypes in this dataset: factors and integers

Methodology

Exploratory data analysis :
The dataset we have used for this project is titled “Human Resources Data Set’’, which has been downloaded from Kaggle (https://www.kaggle.com/datasets/rhuebner/human-resources-dataset?datasetId=1632&searchQuery=method). This dataset has 36 columns and more than 311 values.
We have visualized the data to find the employee categories, Performance salary.. through EDA. Also to Analyse the organization's workforce regarding diversity, performance etc..
By analysing few questions
Is there any relationship between who a person works for and their performance score?
Is there any relation between Performance Score and Special Projects that you worked?
What is the variety of Marital status of the company's employees?
What the best resource to have new employees from it?
We will be analyzing each matter with regards to race, gender and age.
Using the relations it will be easier to predict salaries as they are corelated.

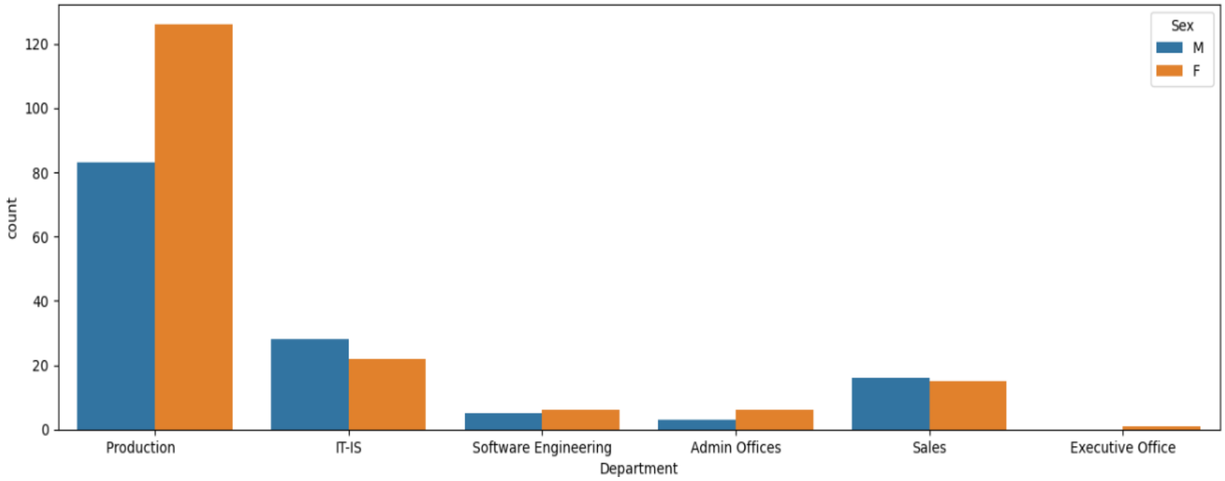
Gender Analysis

```
<AxesSubplot: >
```

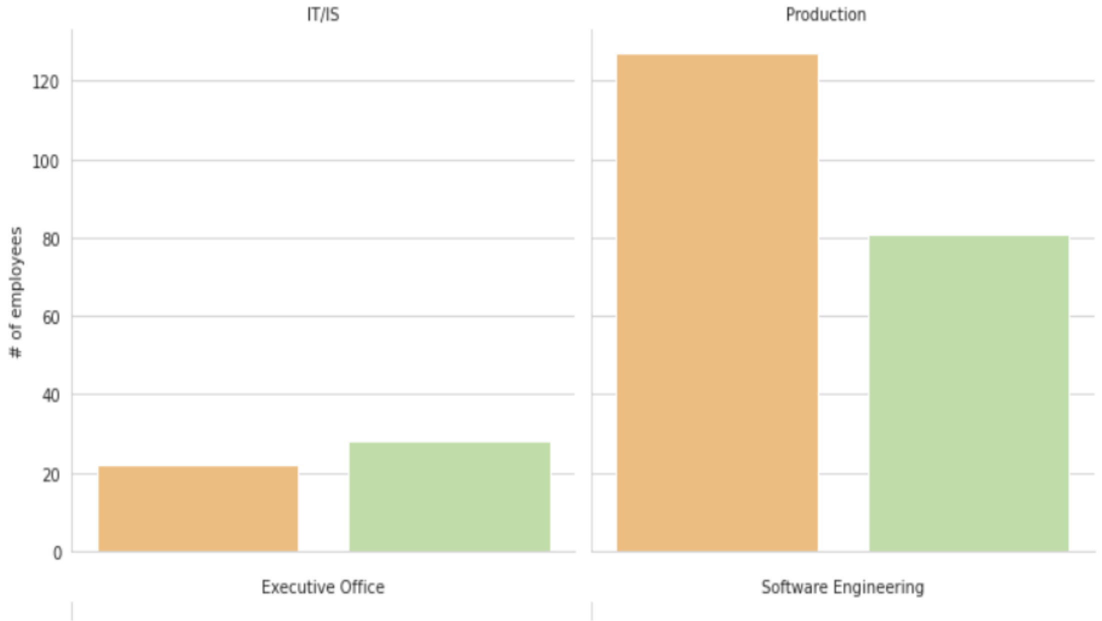


It is clear that the majority of the employees are women. Now we want to know which departments, they work in.

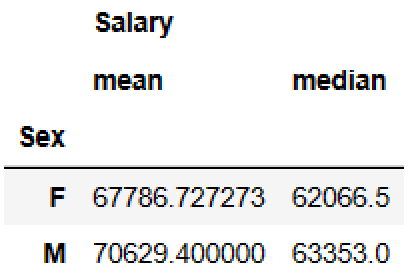
```
<AxesSubplot: xlabel='Department', ylabel='count'>
```



Women work in all departments by close proportion of men, except in production department the women are the majority of the employees.

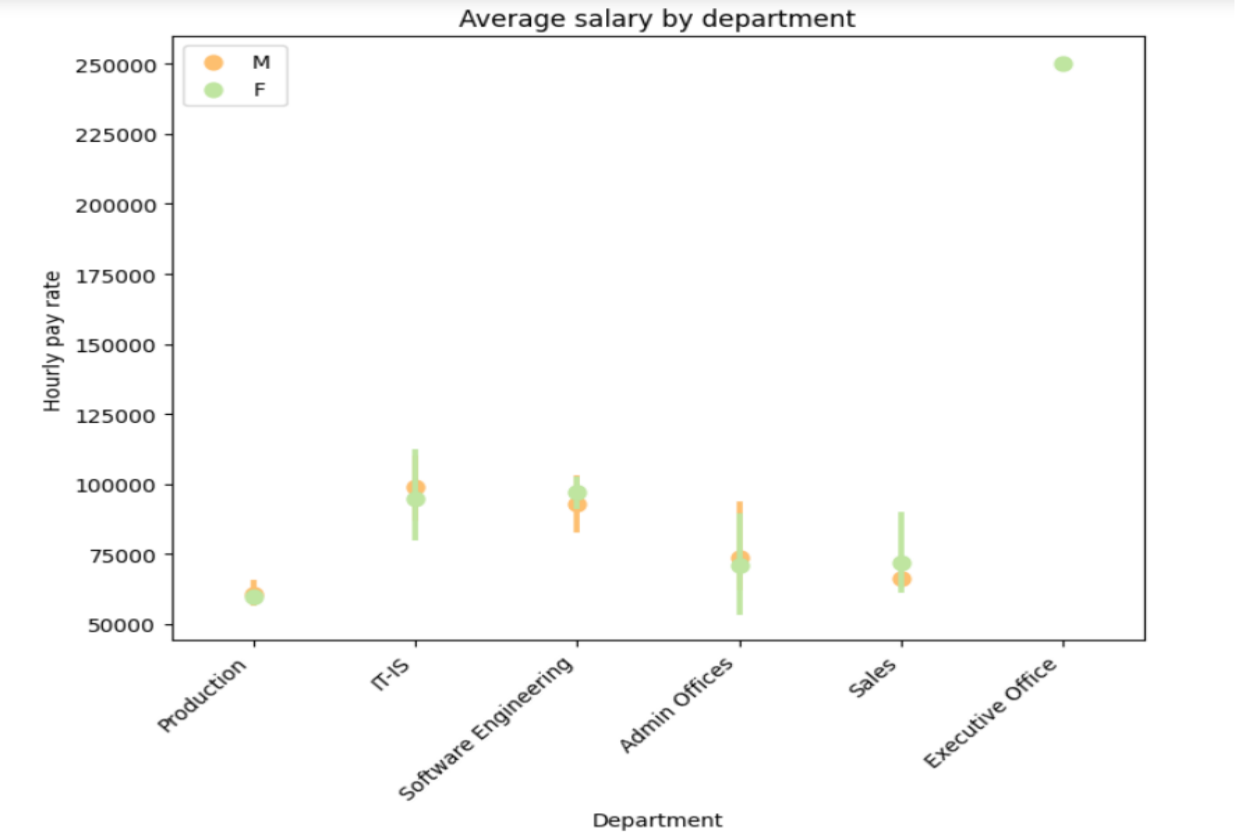


We can see that the numeric difference stems from the *Production department*, where women outnumber men by 50%.



	Salary	
	mean	median
Sex		
F	67786.727273	62066.5
M	70629.400000	63353.0

Women get lower income than men.



At the plot, we identify some divergences in average pay by gender. Women's average pay is slightly higher at Software Engineering department. However, it is important to note that only a handful of employees work there, in such a way that it doesn't affect the overall statistics so much.

On the other hand, Production plays a big part in broadening the wage gap because:

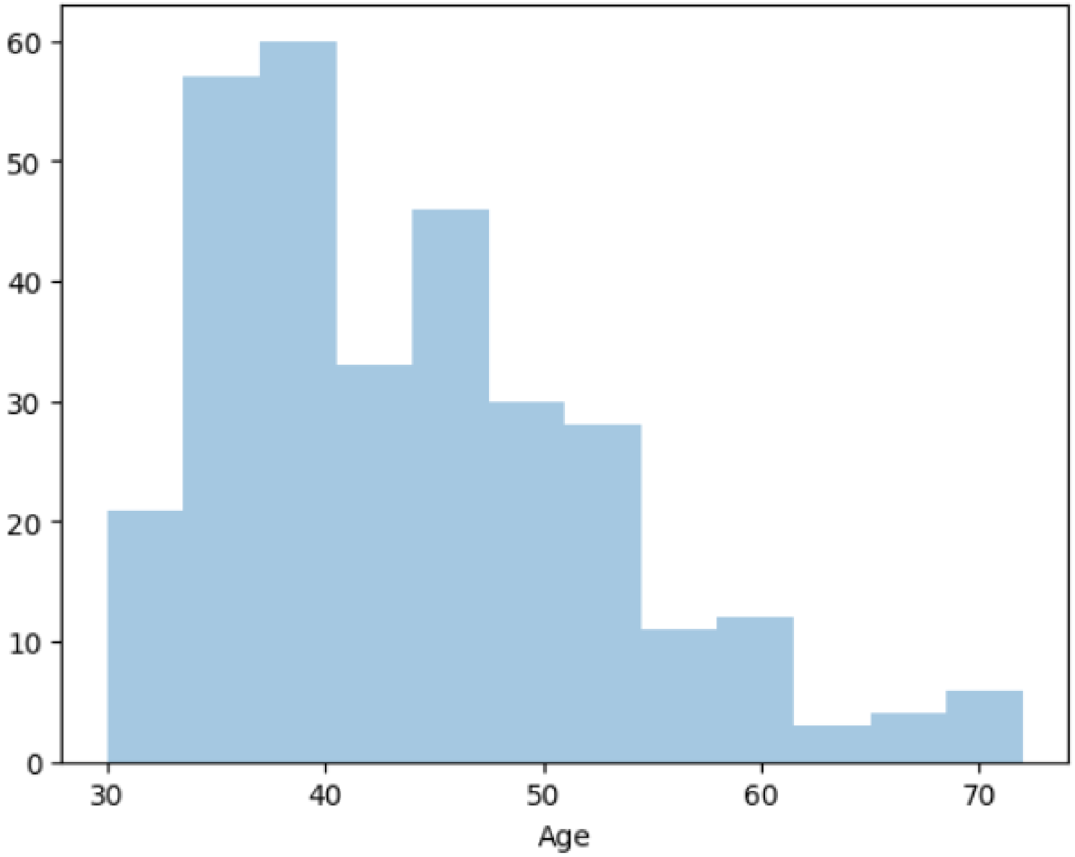
1. the department has the lowest pay rate;

2. this decreases female mean compensation by a lot, since it contains the largest number of workers, most of them being women;

3. considering the employee amount, even though the inequality within Production itself isn't great, it ends up making a significant impact.

We can see a substantial wage gap at Admin Offices; there, women's average income is much lower than men's.

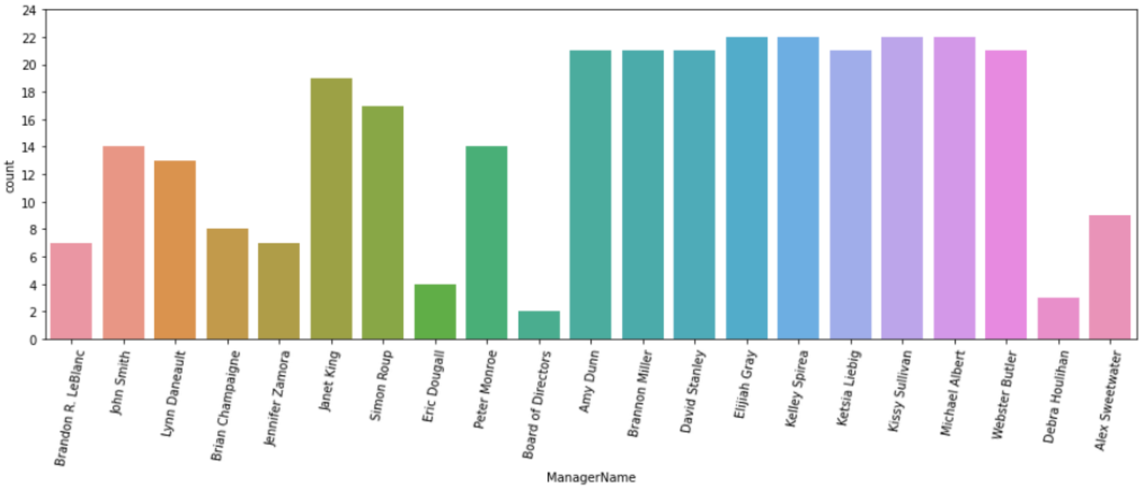
Age Diversity



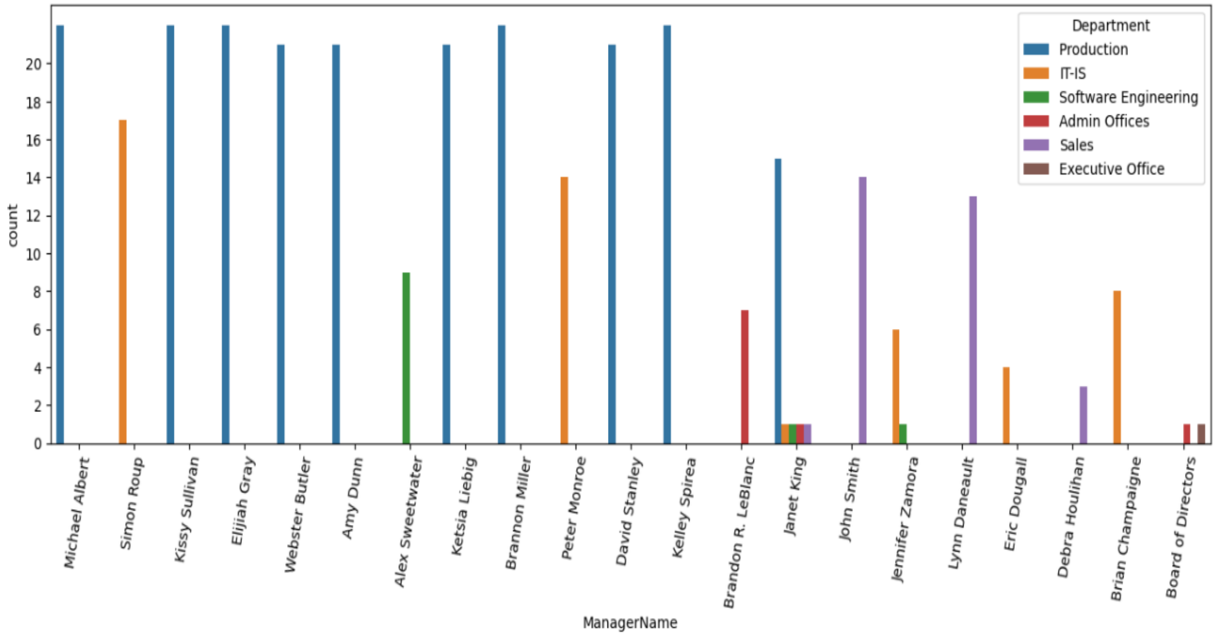
The staff's ages are mainly at early to mid-30s, also counting high at early 40s. The number of employees is substantially lower for ages 55 and over, amounting only to no more than a few dozens, or 6.4%.

Considering how low the numbers go when it comes to older workers, it's important that we inquire how age diversity is promoted through recruiting.

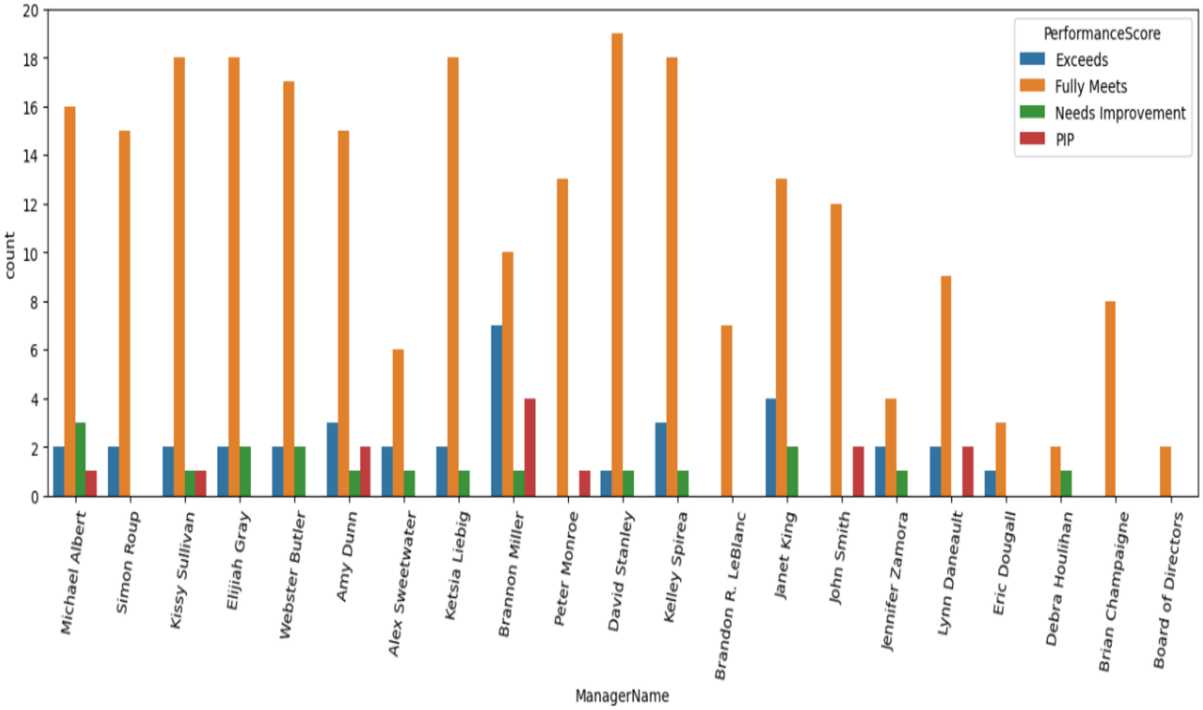
Managers and performance scores:
Is there any relationship between who a person works for and their performance score?



The employees distributed under nearly 14 managers, so what are their departments?



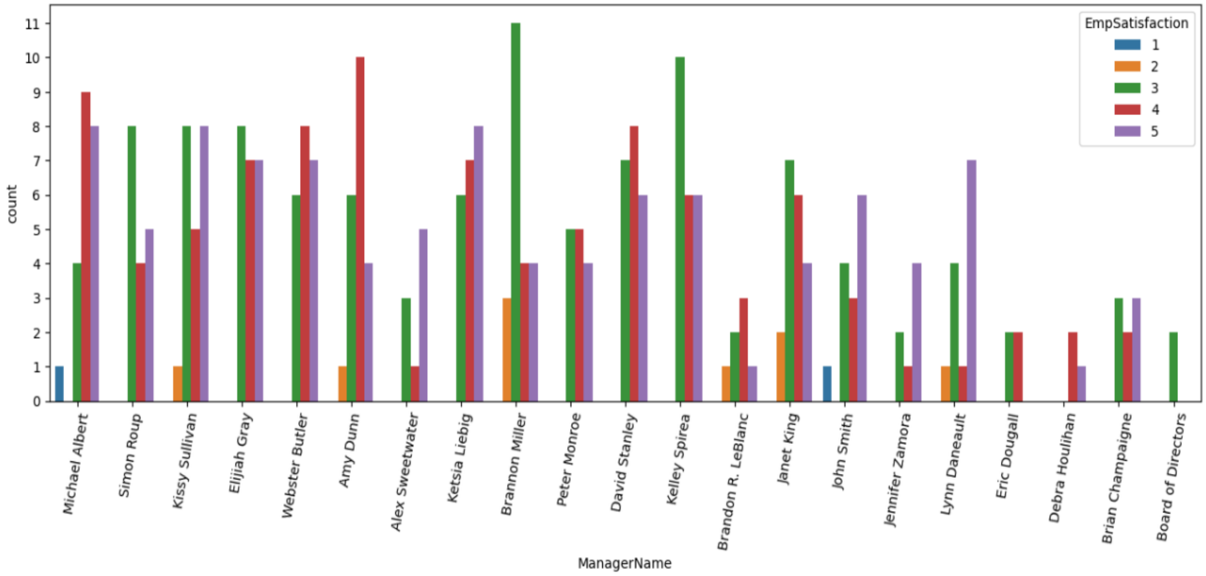
The managers from Production, IT/IS, and Sales. Moreover, most of the employee and managers are in the Production department, following it the IT/IS, then the Sales department comes.



The previous chart is between the manager name and its employees classified on Performance Scores.

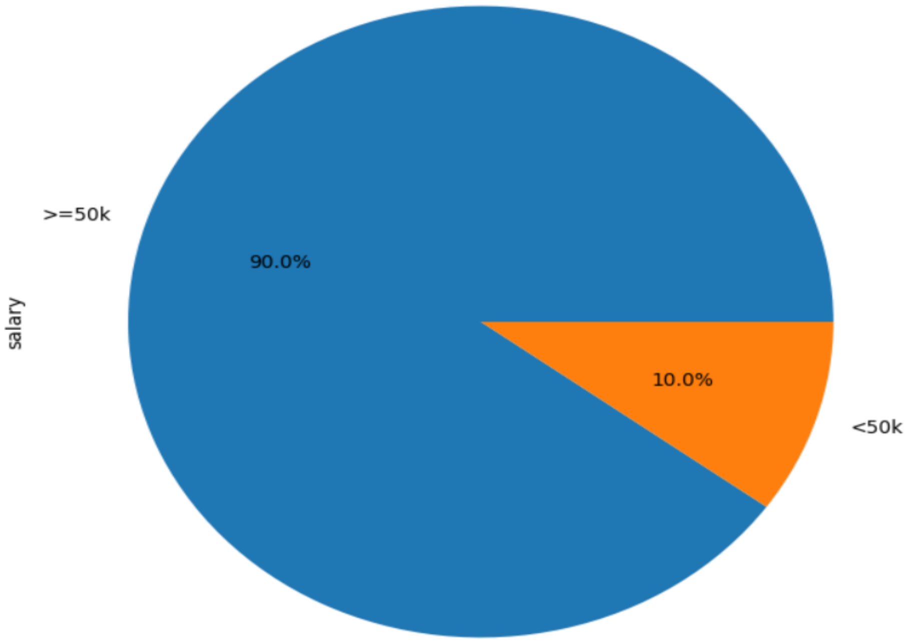
Eor example, The Manager Brannon Miller has the worest Performance in the data frame, whereas David Stanley has the best data set in the dtat frame.

Is there any relation between Employees performances and their employee satisfaction?



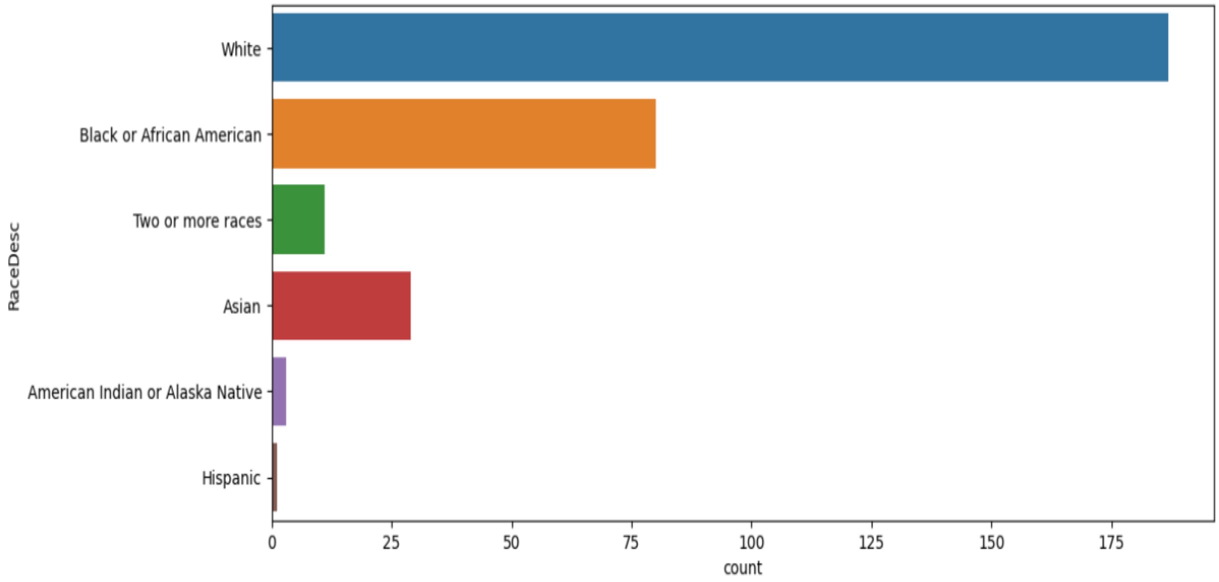
It is noticed that the manager Brannon Miller has the worst employees satisfaction, and this may be because the employees have the worst evaluation in the Preference Scores.

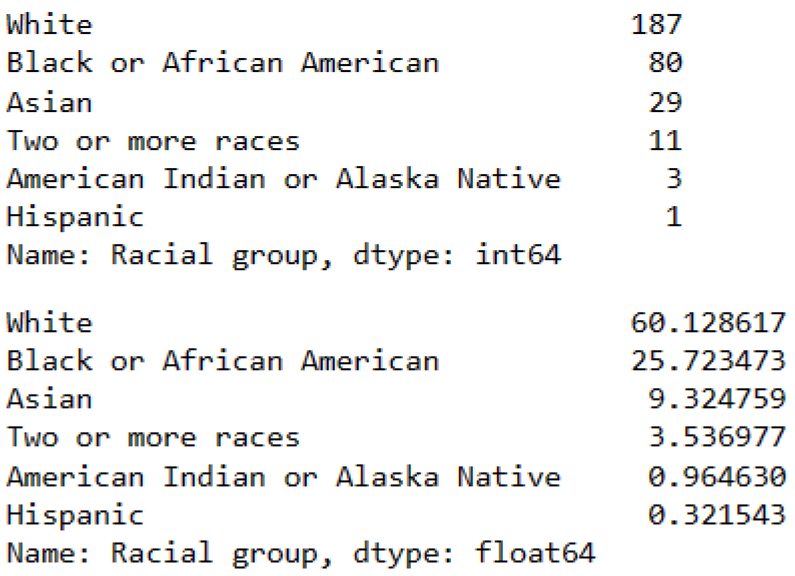
Salary Distribution



Is there discrimination based on race in the company?

```
<AxesSubplot: xlabel='count', ylabel='RaceDesc'>
```

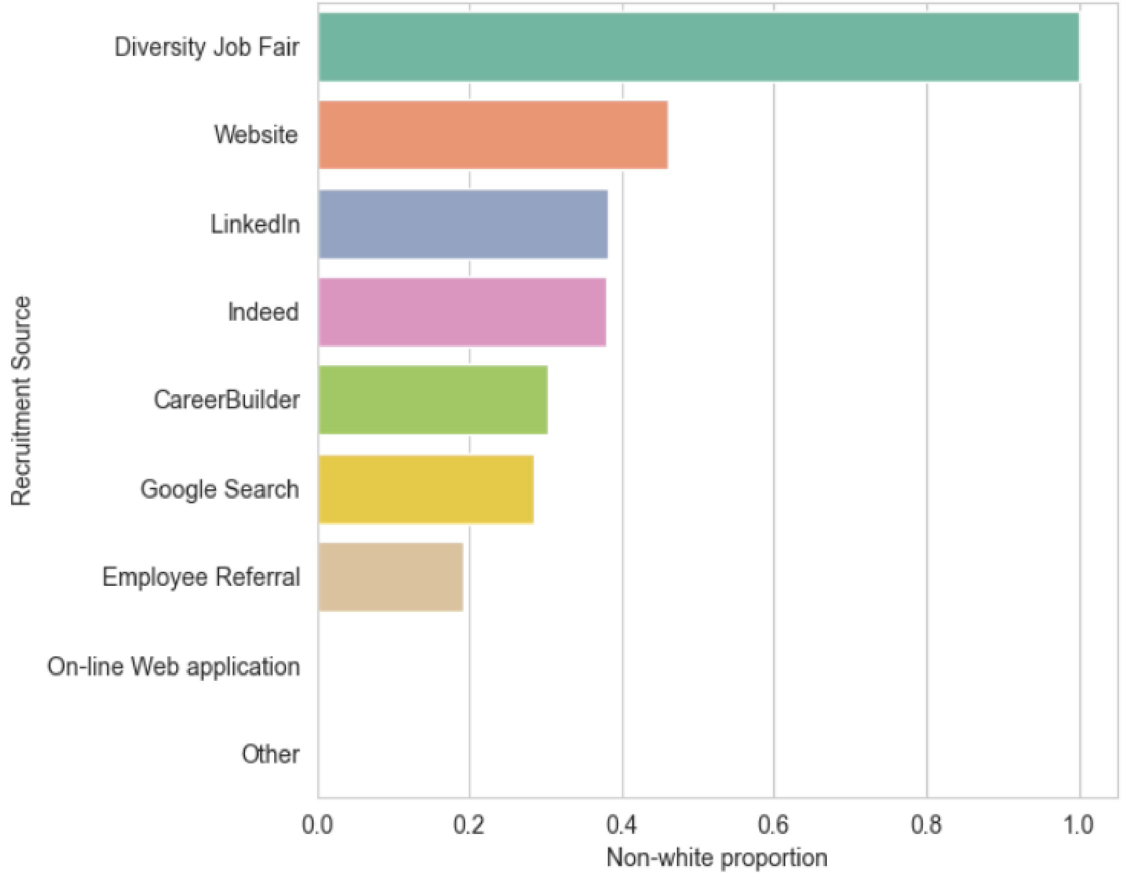




```
White                               187
Black or African American            80
Asian                                29
Two or more races                    11
American Indian or Alaska Native      3
Hispanic                              1
Name: Racial group, dtype: int64

White                               60.128617
Black or African American           25.723473
Asian                                9.324759
Two or more races                    3.536977
American Indian or Alaska Native     0.964630
Hispanic                             0.321543
Name: Racial group, dtype: float64
```

According to 2018 data by the U.S Bureau of Labor Statistics, the country's labor force is made of 78% Whites, 13% Blacks and 6% Asians (https://www.bls.gov/opub/reports/race-and-ethnicity/2018/home.htm). Taking that as a baseline, we can say that the company is diverse from a racial standpoint.



Unsurprisingly, **Diversity Job Fair** plays a crucial role in promoting racial diversity. **Indeed and On-campus Recruiting** can also be lauded for bringing non-white employees more often than not.

On the other hand, **Pay Per Click, On-line Web application, Careerbuilder and Company Intranet** have no contribution to racial diversity at all.

Of course, **respect to diversity includes egalitarian compensation, regardless of skin color.**

Data Preprocessing:

An important step in the process is to make sure that the data is complete and satisfies all the requirements for data analysis. Various pre-processing techniques have been used for the same such as removing invalid rows, changing the column names into more convenient names, performing aggregation on certain columns, converting unix timestamp to proper date format, replacing missing values in columns with the next valid observation, removing unwanted columns and duplicate columns etc. The dataset was age was taken from DOB and data has been processed based on the analysis requirements.

Steps involved in Data Preprocessing:

a. Data Cleaning:

Machine learning algorithms can’t handle missing values and cause problems. So they need to be addressed in the first place. There are many techniques to identify and impute missing values. If a dataset contains missing values and is loaded using pandas, then missing values get replaced with NaN(Not a Number) values. These NaN values can be identified using methods like *isna()* or *isnull()* and they can be imputed using *fillna()*. This process is known as Missing Data Imputation.

The observed row that contains null values has been dropped.

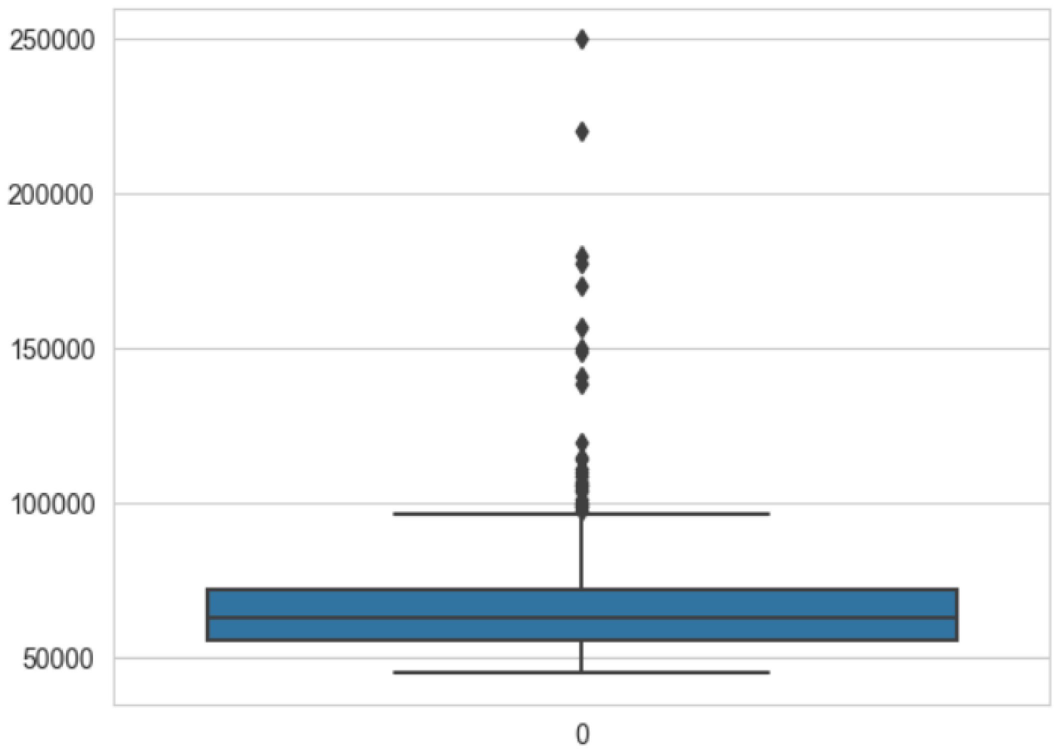
b. Outlier Detection:

An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal)objects. They can be caused by measurement or execution errors. The analysis for outlier detection is referred to as outlier mining. There are many ways to detect the outliers, and the removal process is the data frame same as removing a data item from the panda’s data frame.

An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal)objects. They can be caused by measurement or execution errors. The analysis for outlier detection is referred to as outlier mining. There are many ways to detect the outliers, and the removal process is the data frame same as removing a data item from the panda’s data frame.

Visualizing Outliers Using Box Plot:

It captures the summary of the data effectively and efficiently with only a simple box and whiskers. Boxplot summarizes sample data using 25th, 50th, and 75th percentiles. One can just get insights(quartiles, median, and outliers) into the dataset by just looking at its boxplot.

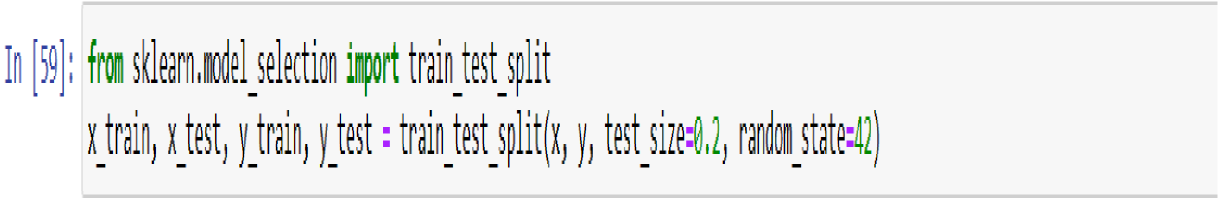


c. Label encoding columns with categorical data

The dataset contained multiple labels in column “icon” and “summary”. To make the data understandable or in human-readable form label encoding was used.Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form.

d. Training and testing the Data

Splitting the dataset into train and test sets is one of the important parts of data pre-processing, as by doing so, we can improve the performance of our model and hence give better predictability. For splitting the dataset, we can use the **train_test_split** function of **scikit-learn.**



```
In [59]: from sklearn.model_selection import train_test_split
         x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

Classification Models

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups.

a. Logistic Regression

It is a supervised learning classification technique that forecasts the likelihood of a target variable. There will only be a choice between two classes. Data can be coded as either one or yes, representing success, or as 0 or no, representing failure. The dependent variable can be predicted most effectively using logistic regression. When the forecast is categorical, such as true or false, yes or no, or a 0 or 1, you can use it. The accuracy of this model is given below.

b. K-Nearest Neighbors

It calculates the likelihood that a data point will join the groups based on which group the data points closest to it are a part of. When using k-NN for classification, you determine how to classify the data according to its nearest neighbor.

c. Support Vector Machine

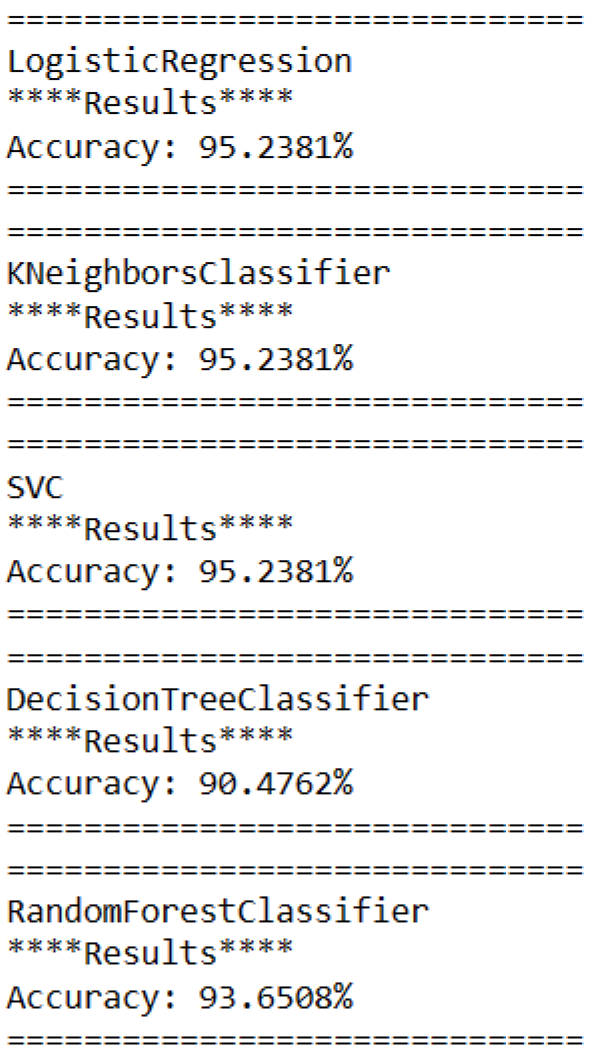
Support Vector Machine is a popular supervised machine learning technique for classification and regression problems. It goes beyond X/Y prediction by using algorithms to classify and train the data according to polarity.

d. Decision Tree

A decision tree is an example of supervised learning. Although it can solve regression and classification problems, it excels in classification problems. Similar to a flow chart, it divides data points into two similar groups at a time, starting with the "tree trunk" and moving through the "branches" and "leaves" until the categories are more closely related to one another.

e. Random Forest Algorithm

The random forest algorithm is an extension of the Decision Tree algorithm where you first create a number of decision trees using training data and then fit your new data into one of the created ‘tree’ as a ‘random forest’. It averages the data to connect it to the nearest tree data based on the data scale. These models are great for improving the decision tree’s problem of forcing data points unnecessarily within a category.

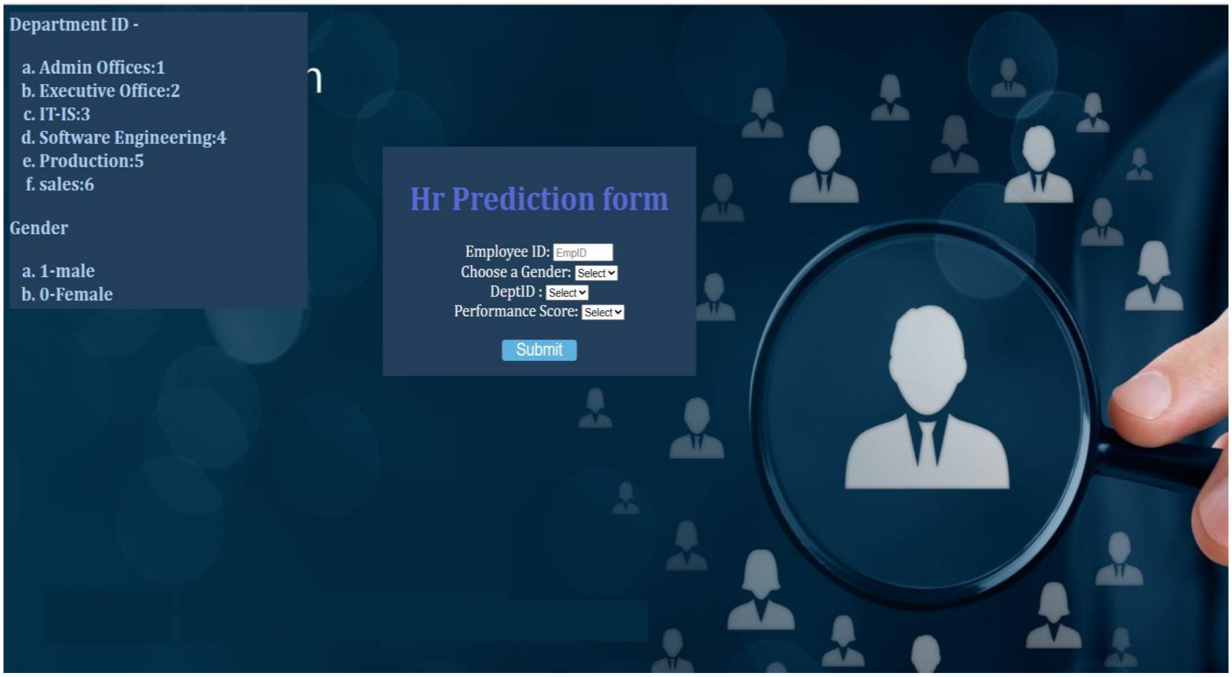


```
==============================
LogisticRegression
****Results****
Accuracy: 95.2381%
==============================
==============================
KNeighborsClassifier
****Results****
Accuracy: 95.2381%
==============================
==============================
SVC
****Results****
Accuracy: 95.2381%
==============================
==============================
DecisionTreeClassifier
****Results****
Accuracy: 90.4762%
==============================
==============================
RandomForestClassifier
****Results****
Accuracy: 93.6508%
==============================
```

Web Hosting:

A web hosting service is a type of Internet hosting service that hosts websites for clients, i.e. it offers the facilities required for them to create and maintain a site and makes it accessible on the World Wide Web. Web page is hosted using Flask and Python anywhere.

Using visual studio code is written that will take care of the server-side processing. Code will receive requests. It will figure out what those requests are dealing with and what they are asking. It will also figure out what response to send to the user. It makes the process of designing a web application simpler. Flask focuses on what the users are requesting and what sort of response to give back. Link is generated t http://127.0.0.1:5000/ in your web-browser, and the webpage as shown below should appear.



Department ID -

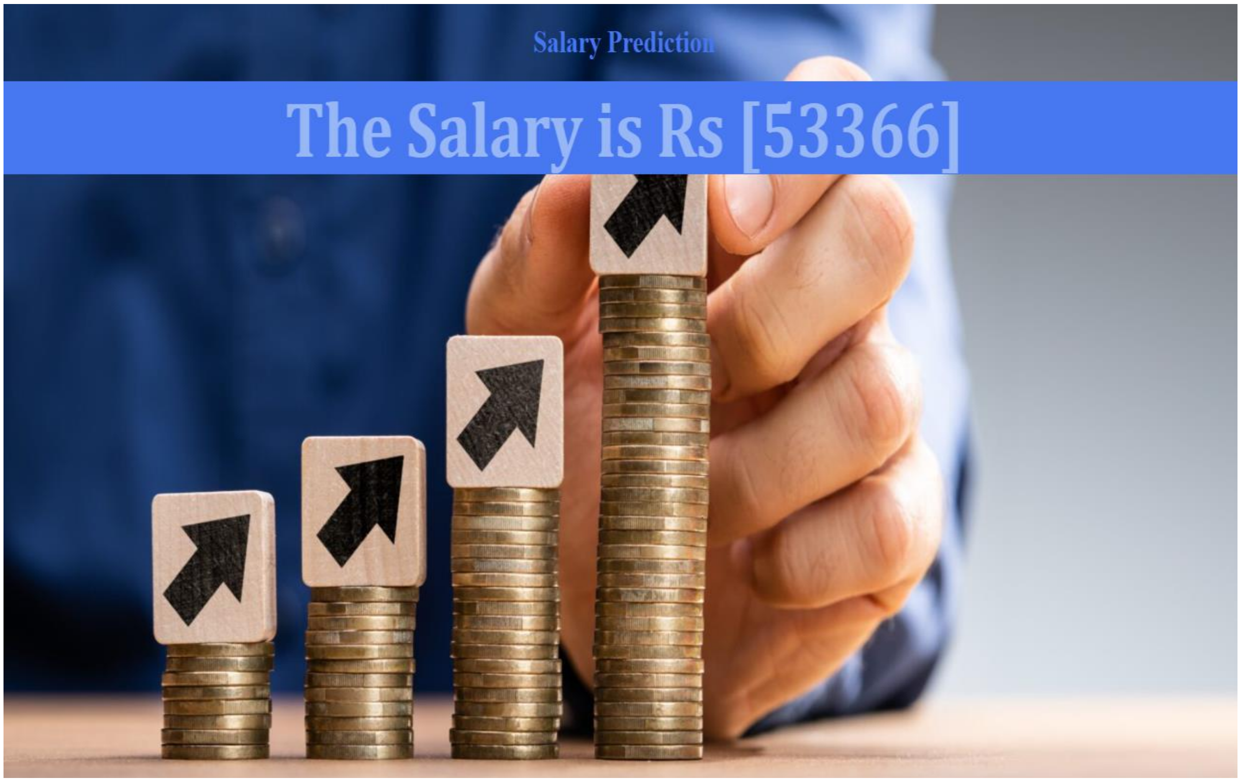
a. Admin Offices:1
b. Executive Office:2
c. IT-IS:3
d. Software Engineering:4
e. Production:5
f. sales:6

Gender

a. 1-male
b. 0-Female

Hr Prediction form

Employee ID: EmpID
Choose a Gender: Select
DeptID : Select
Performance Score: Select

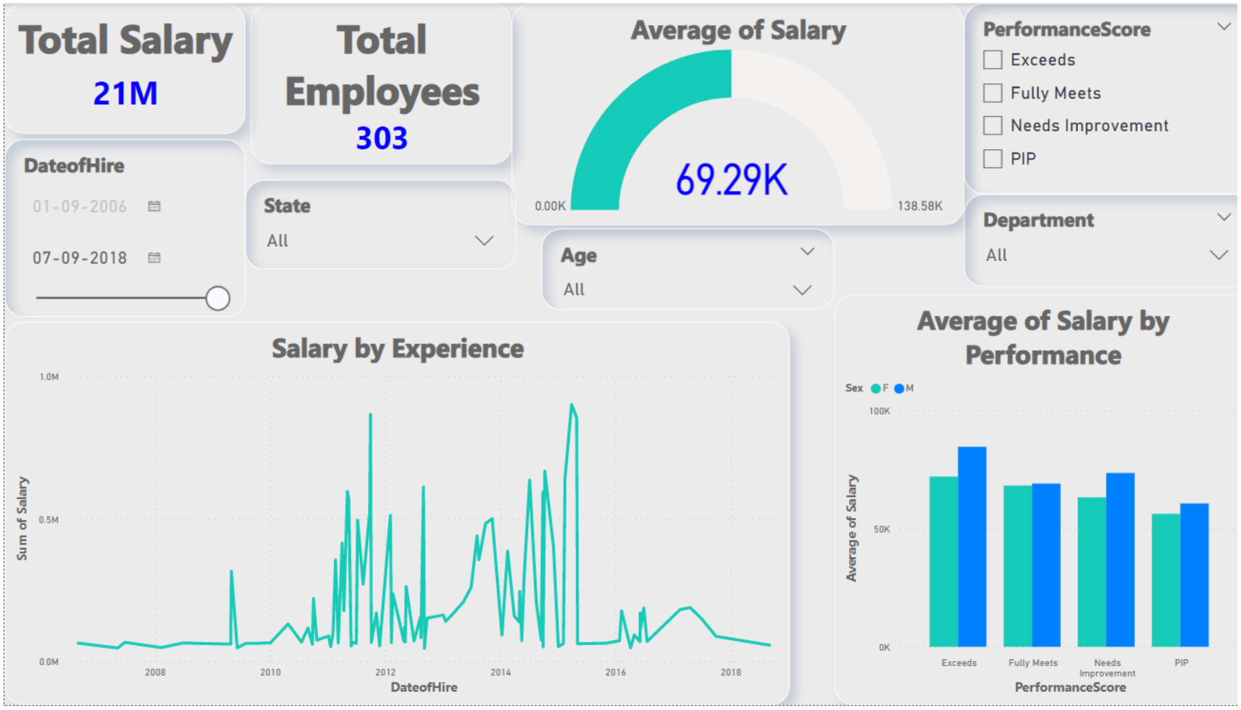


Salary Prediction

The Salary is Rs [53366]

Dashboard in Power BI

A Power BI Dashboard doesn’t only need to be aesthetic but also clear and useful. There are many dimensions by which you can evaluate a Power BI dashboard. The dashboard given below will help you discover useful HR insights for strategic decision making.



Total Salary 21M

Total Employees 303

DateofHire: 01-09-2006 – 07-09-2018

State: All

Age: All

Department: All

PerformanceScore

□ Exceeds

□ Fully Meets

□ Needs Improvement

□ PIP

Findings and Conclusion

1. What are our best recruiting sources if we want to ensure a diverse organization?

It's important that every organization strives to make sure their recruitment practices aren't affected by prejudices or biases. This matter should be addressed with regards to which groups are rejected the most when searching for jobs. **Discrimination is an issue that constantly hits non-white people, women and elders.**

Throughout the analysis of all features, *Diversity Job Fair* proved to be vital in making Dental Magic more plural. It should only be encouraged and expanded.

Regarding race, more than half of the people hired via *Indeed* and *on-campus recruiting* are from underrepresented groups. Contrastingly, some sources need to undergo further scrutiny as to *why they only bring white employees*.

When it comes to age diversity, the organization is far behind, and no recruitment source is distinctly efficient.

2. Are there areas of the company where pay is not equitable?

A deep analysis has showed some wage gaps inside departments:

1. **At *IT/IS*, people of *two or more races* are paid significantly less than other workers in the same job position**. None of the data show a cause for that.
2. **Women's income is lower overall.** The gap stems mainly from two sources:
    - in *Production*, the least well-paid department, females outnumber males by some extent, which results in a greater impact in women's overall salary. Their average income is also slightly inferior in the department;
    - in *Admin Offices*, the wage gap between genders is substantial, though it is hard to tell if the distribution of functions is discriminatory.
3. **Most workers aged 50+ in *Production* work at lower-paid positions.**

3. What is the overall diversity profile of the organization?

Race

While more than half of the workforce is made of white people, we've seen before that the jobs are more well distributed along underrepresented groups comparatively to official statistics in a national level.

Hispanics get a slightly higher pay rate in average, whereas *American Indian or Alaska Natives* perform lower on that variable. Considering that these groups have only four members each, it is premature to conclude the divergence is caused by any discriminatory treatment.

Gender

The workforce is predominantly female. Additionally, Dental Magic's CEO is a woman, and some more can be found in other high positions. These are positive, distinctive traits in a world that favors hiring male workers for most roles, especially leading ones.

A potential highlight is how the company hires many women to work in *Production*, a department where labor is often manual.

However, the company still faces some income inequality related to gender. The issue should be further investigated and dealt with.

Age

Only 6.4% of the workers are 55 or older. This is certainly a diversity issue, specially looking at how 2018 data from the US BLS shows that 23.1% of the workforce in the country is in that age group (https://www.bls.gov/emp/graphics/2019/labor-force-share-by-age-group.htm).

The matter is specially precarious in the Software Engineering department and Admin Offices, that together count no more than a handful employees over ***40***.

Sales, on the other hand, performs really well, having a good amount of well-paid elderly workers.

Link to code and executable file

https://drive.google.com/drive/folders/1-F7C56OSjU_2AUTZSLdyrxYjRfWvAT1G?usp=sharing

Recommendations

It is a generally diverse dataset, though it still has some faults.

The biggest issue is the lack of age diversity, which is not truly promoted by any recruitment source. The organization should review its hiring practices to remove any potential bias.