Rout - DATA 606 Data Project Proposal

Subhalaxmi Rout

Data Preparation

The Company ABC has very poor employee satisfaction and retention. Company has decided to look into the request of paying their employees for overtime hours. The information available for the sample employees includes currently available information such as satisfaction, number of projects and salary level as well as hours worked.

```
# load libraries
library(ggplot2)
library(DT)
library(dplyr)
library(data.table)

# load data
hr_data <- read.csv("https://raw.githubusercontent.com/SubhalaxmiRout002/Data-606-Final-Project/master/

# about data
str(hr_data)
```

```
## 'data.frame':    14999 obs. of  10 variables:
##  $ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
##  $ last_evaluation      : num  0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...
##  $ number_project       : int  2 5 7 5 2 2 6 5 5 2 ...
##  $ average_montly_hours : int  157 262 272 223 159 153 247 259 224 142 ...
##  $ time_spend_company   : int  3 6 4 5 3 3 4 5 5 3 ...
##  $ Work_accident        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ left                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ promotion_last_5years: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Department           : chr  "sales" "sales" "sales" "sales" ...
##  $ salary               : chr  "low" "medium" "medium" "low" ...
```

```
# data info
dim(hr_data)
```

```
## [1] 14999    10
```

```
# view data
DT::datatable(hr_data)
```

Show 10 entries Search:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	Department	salary
1	0.38	0.53	2	157	3	0	1	0	sales	low
2	0.8	0.86	5	262	6	0	1	0	sales	medium
3	0.11	0.88	7	272	4	0	1	0	sales	medium
4	0.72	0.87	5	223	5	0	1	0	sales	low
5	0.37	0.52	2	159	3	0	1	0	sales	low
6	0.41	0.5	2	153	3	0	1	0	sales	low
7	0.1	0.77	6	247	4	0	1	0	sales	low
8	0.92	0.85	5	259	5	0	1	0	sales	low
9	0.89	1	5	224	5	0	1	0	sales	low
10	0.42	0.53	2	142	3	0	1	0	sales	low

Showing 1 to 10 of 14,999 entries Previous 1 2 3 4 5 … 1500 Next

```
#rename columns
colNames <- c("satLevel", "lastEval", "numProj", "avgHrs", "timeCpny"
              , "wrkAcdnt", "left", "fiveYrPrmo", "department", "salary")
data.table::setnames(hr_data, colNames)
names(hr_data)
```

```
##  [1] "satLevel"   "lastEval"   "numProj"    "avgHrs"     "timeCpny"
##  [6] "wrkAcdnt"   "left"       "fiveYrPrmo" "department" "salary"
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

Predict, how much salary the company would need to pay out for the overtime employe.

Cases

What are the cases, and how many are there?

Each case represents an employee working hours details along with salary. There 14999 observations in the given data set.

Data collection

Describe the method of data collection.

Data collected from Kaggle. Here is the source:

Type of study

What type of study is this (observational/experiment)?

This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

Data is collected by `Kaggle` and is available online here: https://www.kaggle.com/giripujar/hr-analytics . For this project, downloaded data from `Kaggle` and stored the data in Github repository. Using `read.csv()` read the data from Git repo.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

Salary, job left are response variable. Both are qualitative.

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

Number of project(number_project), Average Monthly hours spend (average_montly_hours) and level of satisfaction (satisfaction_level) are independent variable. All are quantitative.

Department is qualitative.

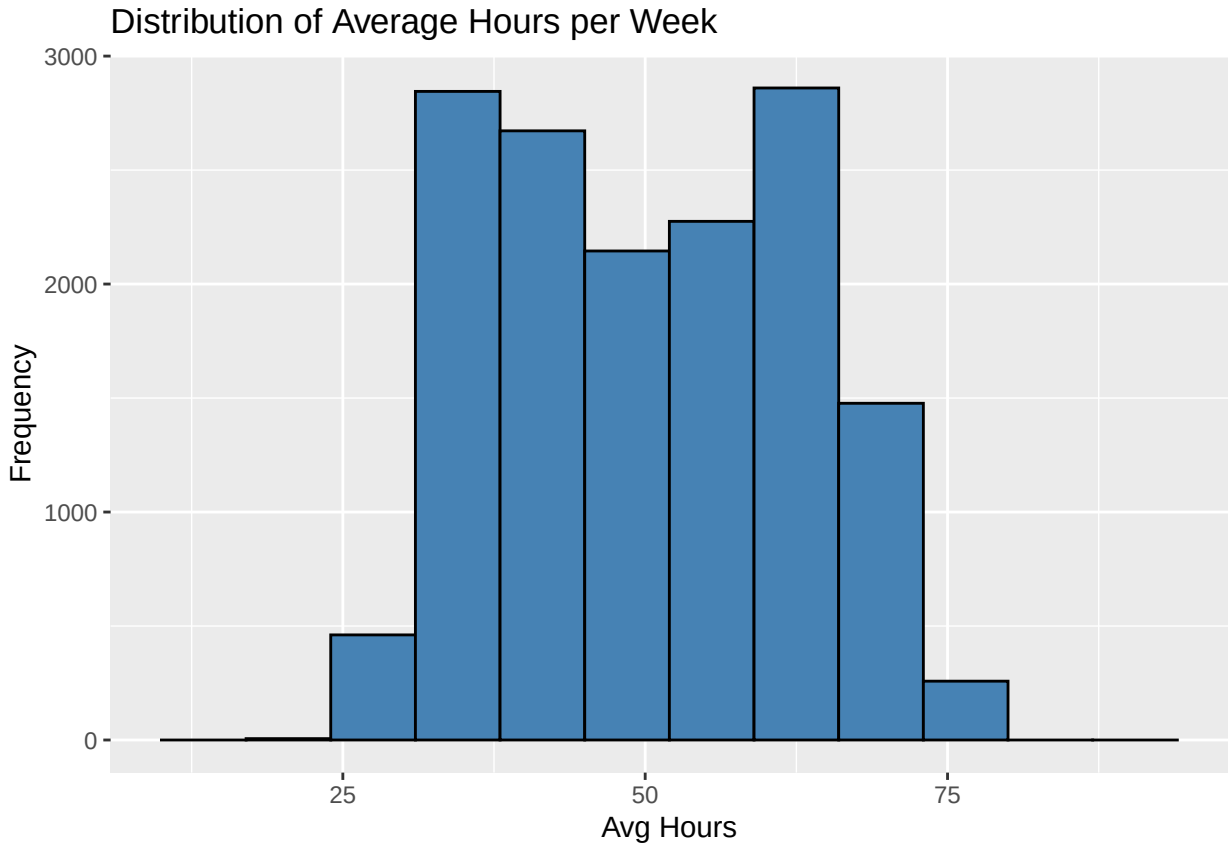
Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

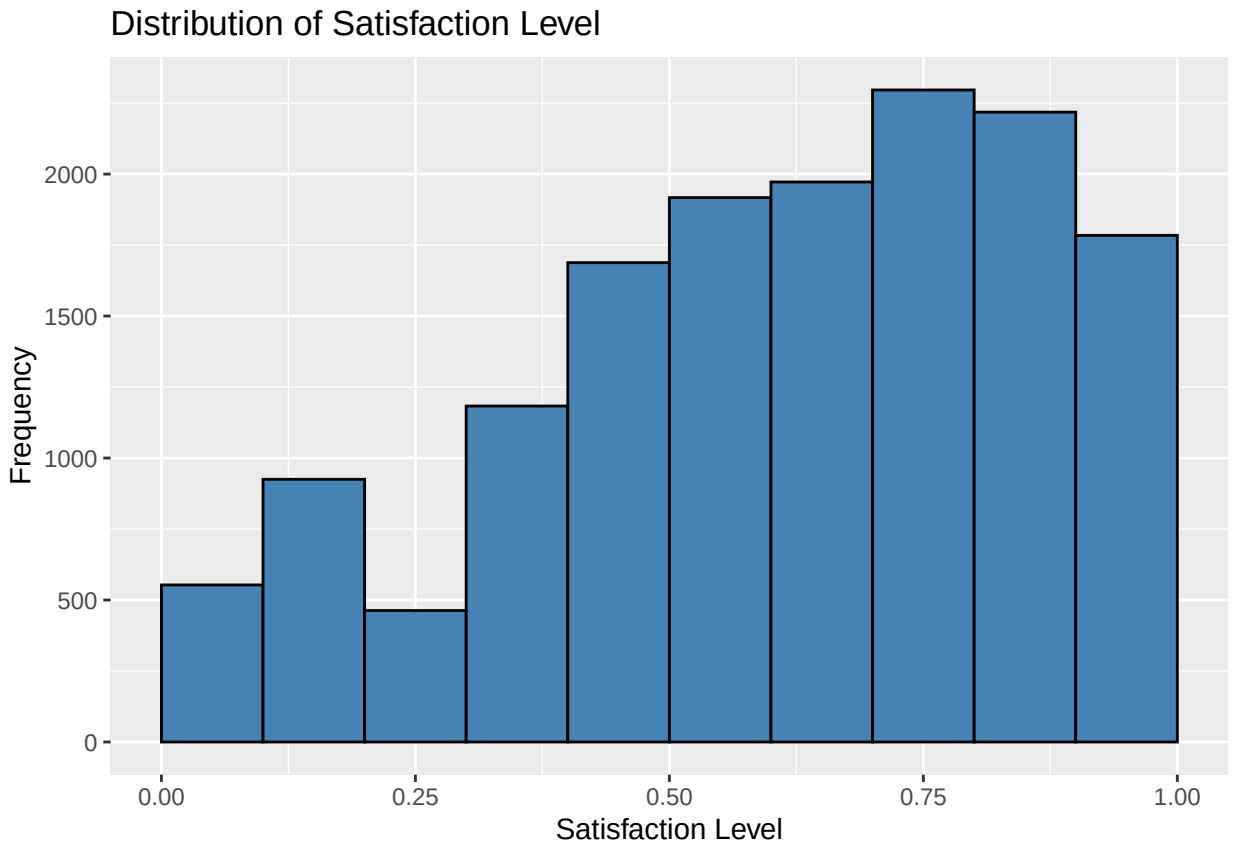
```
# show summary statistics of each column
summary(hr_data)
```

```
##     satLevel         lastEval         numProj          avgHrs
##  Min.   :0.0900   Min.   :0.3600   Min.   :2.000   Min.   : 96.0
##  1st Qu.:0.4400   1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
##  Median :0.6400   Median :0.7200   Median :4.000   Median :200.0
##  Mean   :0.6128   Mean   :0.7161   Mean   :3.803   Mean   :201.1
##  3rd Qu.:0.8200   3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0
##  Max.   :1.0000   Max.   :1.0000   Max.   :7.000   Max.   :310.0
##     timeCpny         wrkAcdnt           left          fiveYrPrmo
##  Min.   : 2.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 3.000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   : 3.498   Mean   :0.1446   Mean   :0.2381   Mean   :0.02127
##  3rd Qu.: 4.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :10.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##   department           salary
##  Length:14999       Length:14999
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

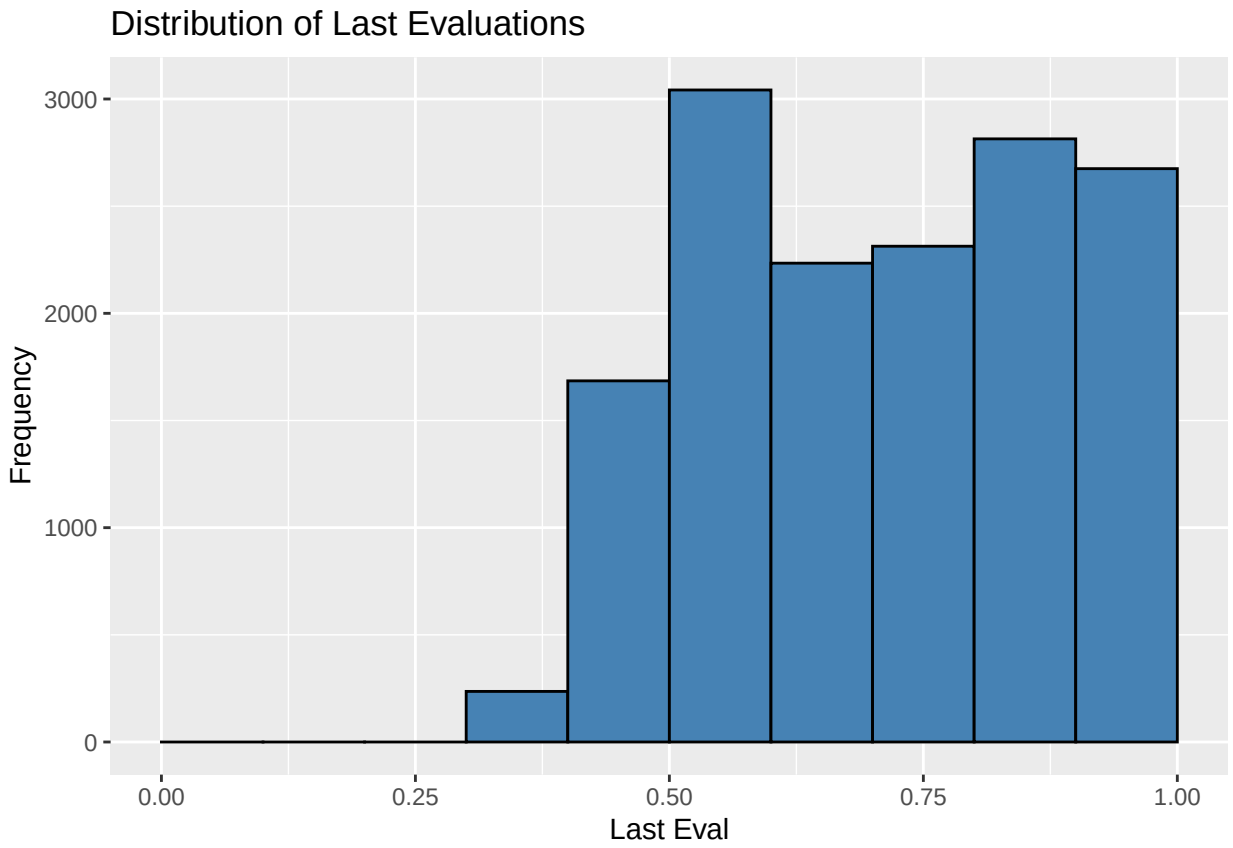
```
# histogram for all numeric variables to understand distribution
ggplot(data = hr_data, aes(x = avgHrs/4)) +
  geom_histogram(breaks=seq(10, 100, by=7), color = "black", fill = "steelblue") +
  labs(title="Distribution of Average Hours per Week",x="Avg Hours", y = "Frequency")
```



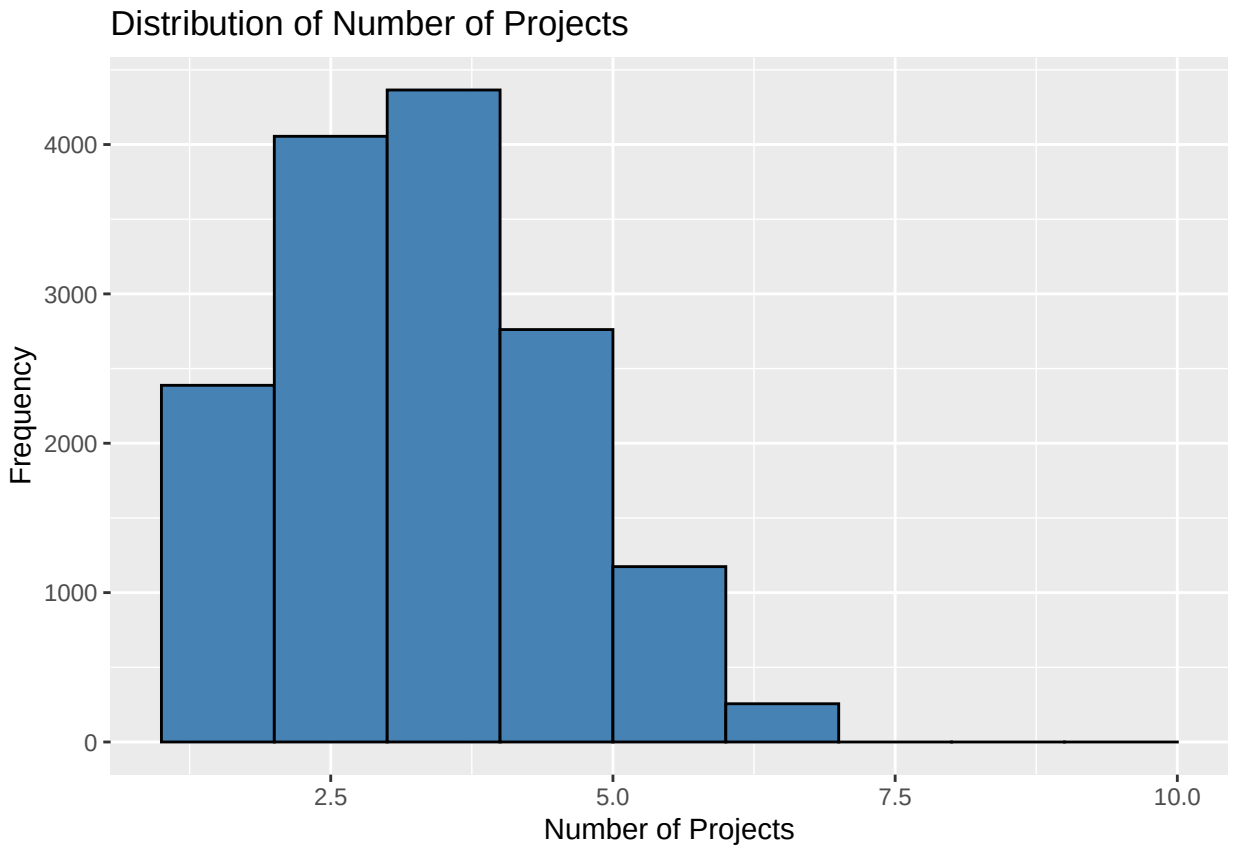

```
ggplot(data = hr_data, aes(x = satLevel)) +
  geom_histogram(breaks=seq(0, 1, by=0.1), color = "black", fill = "steelblue") +
  labs(title="Distribution of Satisfaction Level",x="Satisfaction Level", y = "Frequency")
```




```
ggplot(data = hr_data, aes(x = lastEval)) +
  geom_histogram(breaks=seq(0, 1, by=.1), color = "black", fill = "steelblue") +
  labs(title="Distribution of Last Evaluations",x="Last Eval", y = "Frequency")
```




```
ggplot(data = hr_data, aes(x = numProj)) +
  geom_histogram(breaks=seq(1, 10, by=1), color = "black", fill = "steelblue") +
  labs(title="Distribution of Number of Projects",x="Number of Projects", y = "Frequency")
```




```
# box plot to show the percentile distribution of average hours per week by jdepartment.
ggplot(data = hr_data) + geom_boxplot(aes(x = department, y = avgHrs), fill = "steelblue") +
  labs(title="Department Analysis Data", x="Department", y = "Avg Hours") +
  theme(axis.text.x=element_text(angle=30,hjust=1),plot.title = element_text(hjust = 0.5))
```

Conclusions

We can do a lot of comparions between various variables i.e highest retention by department, employee that decided to left vs the employee that still working based on salary, ratio of satisfied employee vs unsatisfied employee. We can predict the salary (per month/year) for the employee who worked over time using linear regression model.