City Furniture - Employee Turnover
Group 3

Team Members and Roles

Manuel Ucha Rodriguez - Project Lead / Data Scientist

Chenxin Zhang - Data Engineer

Oliver Costantino - Modeler

Lewis Rincon Castano - Data Analyst

Jiayuan Liu - Functionalist

Instructions:

Assignment: You will submit your assignment in Microsoft Word file or Google Docs file to Canvas. Ensure that each of the following components of the presentation are included and best represent your team's assessment:

- A cover page with the team member's names and roles on it. Also include the project name and team number.
- A section with a description of the data sources. (30 points) This section should include:
    - A list of key data tables, including a high level description of how each is anticipated to be used in the project
    - The source of the data tables
    - The acquisition method for the data tables
    - A list of data dictionary files or tables associated with the data tables
    - An identification of important fields or variables contained in the data tables and a description of why they are anticipated to be important in the project
- A section that describes key data manipulation (Data Wrangling) efforts that were performed on the data. (30 points) This section should include:
    - An overall description of the quality of the data:
        - Identify key variables with quality issues
        - Discuss or display the quality issues for key variables by illustrating the issues visually, statistically, or with both methods
        - Explain how the quality issues were addressed
    - Describe reformatting (e.g. converting strings to numeric data) of variables
    - Provide a pseudo-code description of steps taken to clean data for key variables
        - Alternatively, provide links to actual code files that include the steps involved in the cleansing (e.g. SQL, Python, R)
    - Describe how missing data and outliers were addressed (e.g. removal, imputation)
        - Explain why those methods were taken
    - Identify transformations of important data elements that were performed
        - Explain whether the original data was preserved after transformation and why you chose to do so (or not)

- A section that describes the data munging efforts undertaken by the team. (30 Points) This section should include:
    - Identifying additional data sources that were combined in the project to add new variables or data elements
    - Listing the steps taken to merge data from multiple sources
    - Identifying important new fields created from existing data
    - Describing any aggregation methods that were incorporated into the final data tables for modeling
- A section that describes the development workflow for the final data tables that will be used in modeling. (10 points) This section should include the following element:
    - A diagram illustrating data sources and methods used for extracting and manipulating data
        - The diagram should contain references to coding files or notebooks where the data ETL steps were executed (e.g. Python, R, SQL)

Data description

Data Table Name	How it will be used in the project	Source of Data Table	Acquisition Method
ops_turnover_feature_set.xlsx	Data in this table will be used in further modeling and it will be used as the base of the final prediction. We will impute and filter the dataset in order to satisfy the model, which helps us get a result	City Furniture Database	Provided by City Furniture via Email as CSV file
[Pending name]	Employee surveys will be used to analyze their text sentiment and other related methods. These results will work with the ops_turnover_feature_set analysis.	City Furniture Database	Provided by City Furniture via Email as CSV, Excel, or MS Word file

Key variable identification

Key variables	Why is it important?
Age	Based on City Furniture previous analysis, there is a bigger chance of their drivers leaving within the first year of work. We believe age could be an indicator to predict the turnover
Weekend_Shifts_Per_Week_Last_Month_Average	Our team predicted “Weekend_Shifts_Per_Week_Last_Month_Average” to be an important indicator of turnover since employees usually want their weekends off to enjoy free time.
Shifts_per_week_last_month_average	We brainstormed “Hours_worked_per_week_last_month_average” to be significant since employees with a high number of shifts worked last month when the data was collected could be more likely to leave given their high number of shifts worked.

Hours_worked_per_week_last_month_avgerage	We brainstormed “Hours_worked_per_week_last_month_avgerage” to be significant since employees with a high number of hours worked last month when the data was collected could be more likely to leave given their high number of hours worked.
Returns_90_Days	“Returns_90_days” could be a good indicator since a high number of returns leading up to data collection could correlate with employee turnover
Overnight_Shifts	We consider drivers would rather work in the daytime than the nighttime which means this variable could be correlated to the turnover rate.

Data Dictionary

ops_turnover-Data Dictionary.png

	FEATURE	FEATURE SET DEFINITIONS
0	CAREER_PEICES_DELIVERED	Number of pieces delivered by each tech durin their career
1	CAREER_DAILY_DELIVERED_AVERAGE	Average number of daily delivered pieces by each tech during their career
2	CAREER_6_MONTH_DELIVERED_AVERAGE	Average number of daily delivered pieces by each tech during past 6 months
3	Daily Deliveries Past Month Average	Average number of daily delivered pieces by each tech during past month
4	PEICES_DELIVERED_90_DAYS	Total number of pieces delivered in the past 90 days
5	CAREER_RETURNS	Total number of pieces returned in the past 90 days
6	CAREER_DAILY_RETURN_AVERAGE	Average number of daily return pieces over total career
7	SIX_MONTH_DAILY_RETURN_AVERAGE	Average number of daily return pieces over 6 month period
8	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	Average number of daily return pieces over past month
9	RETURNS_90_DAYS	Total number of pieces return in the past 90 days
10	CAREER_SHIFTS	Total number of shifts worked by associate
11	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	Average number of shifts over the past 6 months per week
12	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	Average number of shifts over the past month
13	OVERNIGHT_SHIFTS	Total number of overnight shifts worked by associate
14	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	Average number of overnight shifts over past 6 months
15	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	Average number of of overnight shifts over past month
16	WEEKEND_SHIFTS	Total number of overnight shifts worked by assocaite
17	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	Average number of weekend shifts over past 6 months
18	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	Average number of weekend shifts over past month
19	WEEK_SHIFTS	Total number of shifts worked WTD
20	STATUS	Current Employee Statues
21	CAREER_HOURS_WORKED	Total number of hours worked over employee career
22	CAREER_HOURS_WORKED_DAILY_AVERAGE	Average number of hours worked per shift over career
23	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	Average number of hours worked per shift over last 6 months
24	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	Average number of hours worked per shift over last month
25	TECH	Employee Identification number
26	DELETE_CODE	Code identifying employee delete status
27	AGE	The current age of the employee
28	STATUS	0 or 1 where 0 is active employee and 1 is terminated/loa employee
29	JOB_DESCRIPTION	Description of position

Data Type

```
> sapply(df_cf, class)
                                      ...1                                       TECH
                                  "numeric"                                  "numeric"
                                DELETE_CODE                                        AGE
                                 "character"                                  "numeric"
                                    STATUS                            JOB_DESCRIPTION
                                  "numeric"                                "character"
  WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE                          CAREER_SHIFTS
                                  "numeric"                                  "numeric"
     HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE                         WEEKEND_SHIFTS
                                  "numeric"                                  "numeric"
           SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE          CAREER_6_MONTH_DELIVERED_AVERAGE
                                  "numeric"                                  "numeric"
    HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE                         CAREER_RETURNS
                                  "numeric"                                  "numeric"
          SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE              CAREER_DAILY_RETURN_AVERAGE
                                  "numeric"                                  "numeric"
                              RETURNS_90_DAYS                 PEICES_DELIVERED_90_DAYS
                                  "numeric"                                  "numeric"
                    CAREER_PEICES_DELIVERED                       CAREER_HOURS_WORKED
                                  "numeric"                                  "numeric"
          CAREER_HOURS_WORKED_DAILY_AVERAGE          SIX_MONTH_DAILY_RETURN_AVERAGE
                                  "numeric"                                  "numeric"
                               WEEK_SHIFTS  PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE
                                  "numeric"                                  "numeric"
             CAREER_DAILY_DELIVERED_AVERAGE                        OVERNIGHT_SHIFTS
                                  "numeric"                                  "numeric"
          DAILY_DELIVERY_PAST_MONTH_AVERAGE OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
                                  "numeric"                                  "numeric"
OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE  WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
                                  "numeric"                                  "numeric"
```

Data Wrangling

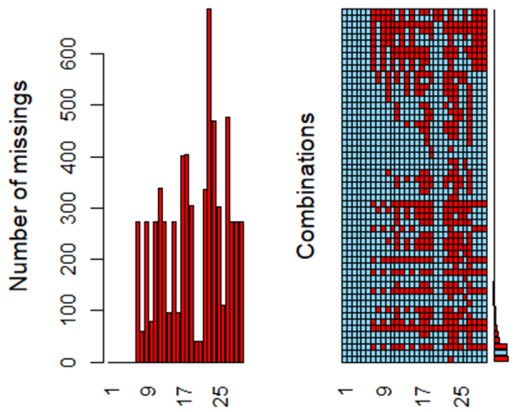
Data wrangling - Missing Values

- No data has been removed at this moment.

- Below we included the information related to the missing values from the "ops_turnover_feature_set" file.
- On Decemeber 8th meeting with City Furniture team, the Student Team was instructed to do not remove the outliers from the dataset.

Coding	Output

```r
# Load libraries and import dataset
library(readxl)
library(mice) # Imputation method with Mice
library(Hmisc) # Imputation with mean
library(tidyverse)
library(dplyr)
library(VIM) # Vizualization for missing data
library("writexl") # To export our imputed tables as an Excel tabl

ops_turnover_feature_set <- read_excel("ops_turnover_feature_set.x
View(ops_turnover_feature_set)
df_cf <- ops_turnover_feature_set
summary(df_cf)

# Find missing values and create graph
aggr(df_cf, prop=FALSE, numbers=TRUE)
```




```r
# Percentage of missing values by column
colSums(is.na(df_cf)) * 100 / nrow(df_cf)
# Number of missing values per column
colSums(is.na(df_cf))
```

- From this output, we found only six columns without missing values.
- The range of missing values is from 30% to almost 77%.
  - For instance, the variable with this 77% is "week_shifts"
  - The remaining variables have a range of missing values between 30% and 53%.
  - The student team asked about these missing values to the City Furniture team. They replied that the data obtained to build this table is inconsistent since only some employees report the same amount of information.
  - City Furniture reported that this lack of information is due to multiple reasons, such as the employee being recently hired or not providing enough information.

We have more missing values on the following columns and their index number:

Percentage of Missing Values

```
> colSums(is.na(df_cf)) * 100 / nrow(df_cf)
                                        ...1
                                    0.000000
                                 DELETE_CODE
                                    0.000000
                                      STATUS
                                    0.000000
  WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
                                   30.536913
     HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE
                                   30.536913
           SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
                                   30.536913
    HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE
                                   30.536913
          SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
                                   30.536913
                              RETURNS_90_DAYS
                                   44.966443
                      CAREER_PEICES_DELIVERED
                                   34.004474
           CAREER_HOURS_WORKED_DAILY_AVERAGE
                                    4.362416
                                 WEEK_SHIFTS
                                   76.845638
              CAREER_DAILY_DELIVERED_AVERAGE
                                   33.668904
           DAILY_DELIVERY_PAST_MONTH_AVERAGE
                                   53.243848
OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
                                   30.536913
```

- Week_shifts (23) with 687 missing values
- Daily Delivery Past Month Average (27) with 476 missing values
- Past Month Pieces Return Daily Average (24) with 470 missing values

To replace our missing values, City furniture was adamant that we should interpolate these values rather than deleting since our model needed all the data possible

```
                                                TECH
                                            0.000000
                                                 AGE
                                            0.000000
                                      JOB_DESCRIPTION
                                            0.000000
                                      CAREER_SHIFTS
                                            6.599553
                                      WEEKEND_SHIFTS
                                            8.836689
                  CAREER_6_MONTH_DELIVERED_AVERAGE
                                           37.919463
                                      CAREER_RETURNS
                                           10.850112
                        CAREER_DAILY_RETURN_AVERAGE
                                           10.850112
                         PEICES_DELIVERED_90_DAYS
                                           45.190157
                              CAREER_HOURS_WORKED
                                            4.362416
                   SIX_MONTH_DAILY_RETURN_AVERAGE
                                           37.695749
          PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE
                                           52.572707
                                OVERNIGHT_SHIFTS
                                           12.304251
OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
                                           30.536913
  WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
                                           30.536913
```

Missing Values Per Column

```
> colSums(is.na(df_cf))
                                                ...1
                                                   0
                                         DELETE_CODE
                                                   0
                                              STATUS
                                                   0
  WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
                                                 273
     HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE
                                                 273
           SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
                                                 273
    HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE
                                                 273
          SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
                                                 273
                                RETURNS_90_DAYS
                                                 402
                     CAREER_PEICES_DELIVERED
                                                 304
           CAREER_HOURS_WORKED_DAILY_AVERAGE
                                                  39
                                   WEEK_SHIFTS
                                                 687
              CAREER_DAILY_DELIVERED_AVERAGE
                                                 301
           DAILY_DELIVERY_PAST_MONTH_AVERAGE
                                                 476
OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
                                                 273
```

	TECH 0 AGE 0 JOB_DESCRIPTION 0 CAREER_SHIFTS 59 WEEKEND_SHIFTS 79 CAREER_6_MONTH_DELIVERED_AVERAGE 339 CAREER_RETURNS 97 CAREER_DAILY_RETURN_AVERAGE 97 PEICES_DELIVERED_90_DAYS 404 CAREER_HOURS_WORKED 39 SIX_MONTH_DAILY_RETURN_AVERAGE 337 PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE 470 OVERNIGHT_SHIFTS 110 OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE 273 WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE 273
R file:	**Assignment_5_boxplots.R**

- **Outliers:** we are waiting to receive the pending employees' survey data to start considering any future removal of values or not.

```
41 ```{r}
42 # List of the columns names and their index number
43 df_cf2 <- colnames(ops_turnover_feature_set)
44 columnNames <- as.data.frame(df_cf2)
45 ```
```

	df_cf2
1	...1
2	TECH
3	DELETE_CODE
4	AGE
5	STATUS
6	JOB_DESCRIPTION
7	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
8	CAREER_SHIFTS
9	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE
10	WEEKEND_SHIFTS
11	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
12	CAREER_6_MONTH_DELIVERED_AVERAGE

	df_cf2
13	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE
14	CAREER_RETURNS
15	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
16	CAREER_DAILY_RETURN_AVERAGE
17	RETURNS_90_DAYS
18	PEICES_DELIVERED_90_DAYS
19	CAREER_PEICES_DELIVERED
20	CAREER_HOURS_WORKED
21	CAREER_HOURS_WORKED_DAILY_AVERAGE
22	SIX_MONTH_DAILY_RETURN_AVERAGE
23	WEEK_SHIFTS
24	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE

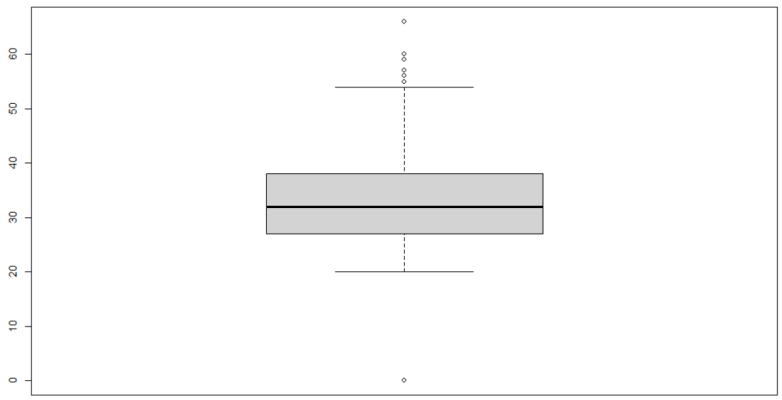
25	CAREER_DAILY_DELIVERED_AVERAGE
26	OVERNIGHT_SHIFTS
27	DAILY_DELIVERY_PAST_MONTH_AVERAGE
28	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
29	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
30	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE

Coding

```
# Create boxplots with single variable
boxplot(ops_turnover_feature_set$AGE,
        xlab='AGE')
boxplot(ops_turnover_feature_set$CAREER_R
        xlab='Career Returns')
# Create boxplots with multiple variables
colnames(df_cf) <- c(1:30)
boxplot(df_cf[7:12])
boxplot(df_cf[13:18])
boxplot(df_cf[19:24])
boxplot(df_cf[25:30])
```

- For better visualization, the columns were renamed with numbers from 1 to 30.
- Boxplots started from column seven because the five of the six columns do not have missing values or are characters observations.
- There are only six variables with not outliers:

Output

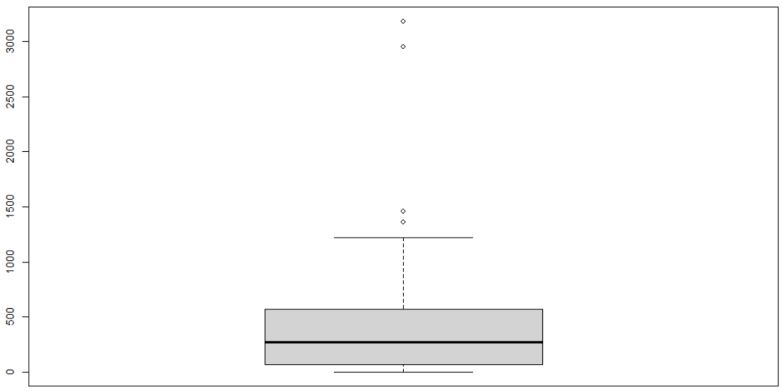


Age

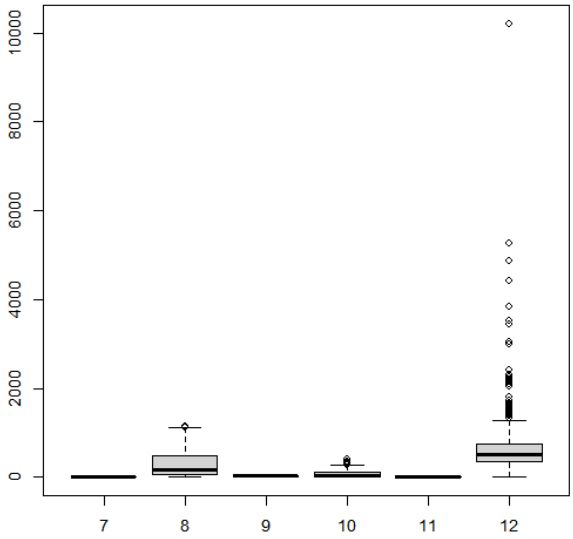
- Weekend_shifts_per_week_last_month_average (7)
- Hours_worked_per_week_six_month_averag (9)
- Shifts_per_week_six_month_average (11)
- Hours_worked_per_week_last_month_average (13)
- Shifts_per_week_last_month_average (15)
- Weekend_shifts_per_week_six_month_average (30).

- In regards to our key variables, we identified that:
  - Age has some outliers. For instance, we have one employee with age zero. However, the highest values that are over 60 should not be a concern since retirement age is around 60-70s in most cases.
  - Weekend_shifts (#16) and overnight_shift (#26) have a few outliers but are not a concern because this variable is taking into consideration the work record of new and senior employees.
  - Career_returns (#14) has a few outliers.

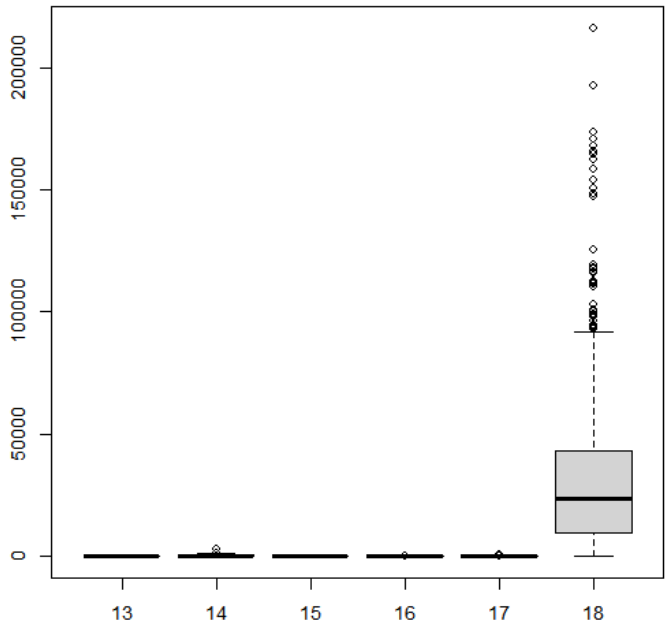
City furniture concluded that we should build a model

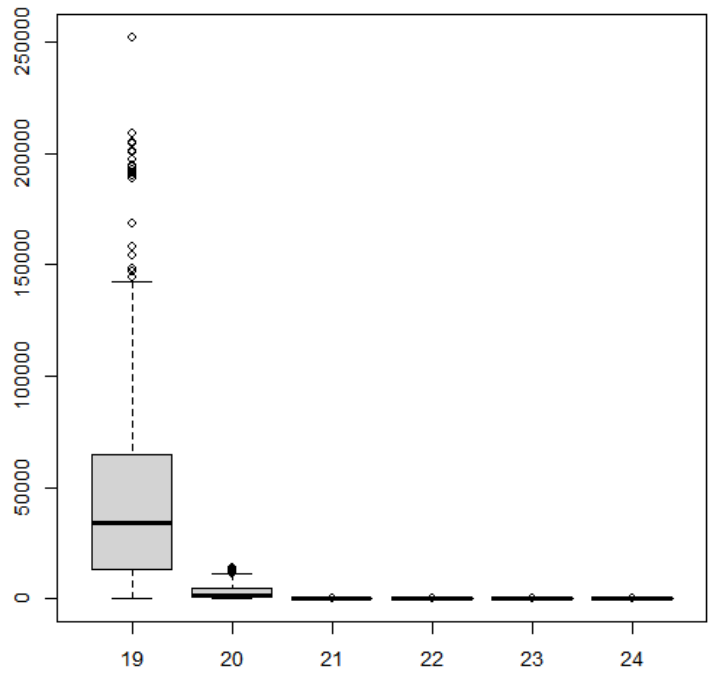


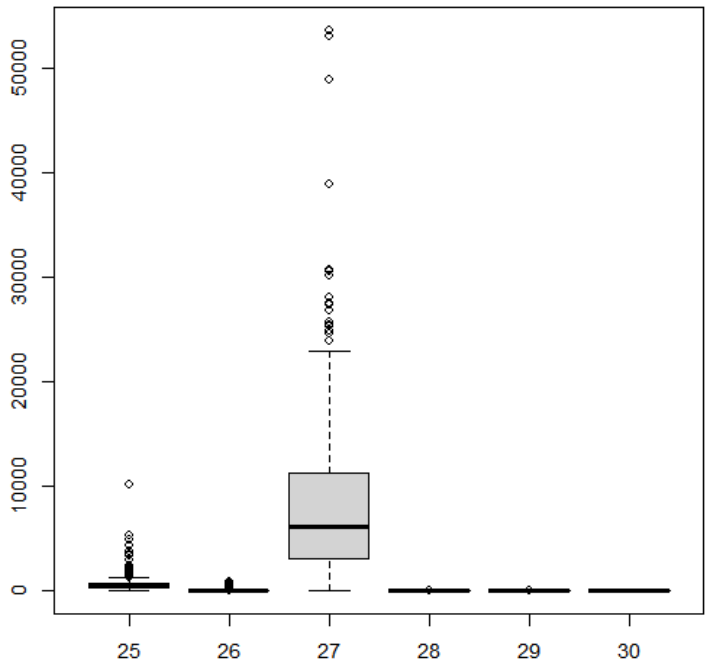
Career Returns



including the outliers and a second model with the outliers removed.





	
R File:	Assignment_5_outliers.R

Imputation Methods

Since we have missing values, we are required to fill these values in order to have a better table before data modeling. During our whiteboard with our class TA, Tek Song, he recommended us to create five imputations. Then, we will compare them later during our modeling process

Mice - Predictive Mean Matching (PMM) Method - Coding	
• As we can observe in the below table, we do not have missing values after performing the mean	29 # Apply Mice method for missing values 30 # Source: 31 # https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/ 32 imputed_Data <- mice(df_cf, m=5, maxit = 50, method = 'pmm', seed = 500) 33 completeData_mice_method <- complete(imputed_Data,3) 34 anyNA(completeData_mice_method) # Returns false for missing values 35 summary(completeData_mice_method) 36

imputation.

- New imputation file was uploaded to OneDrive.

```
# Export dataframe as a file on 11/30/2022
write_xlsx(completeData_mice_method,
           "C:\\Users\\lewis\\Desktop\\completeData_mice_method.xlsx")
```

Missing Value? No, the table below has proof of this absence of NaN values.

```
# Information about the PMM dataframe
colnum <- (seq_len(ncol(completeData_mice_PMM_method)))
PMM_colinfo <- as.data.frame(colnum)
PMM_colinfo$variable <- names(completeData_mice_PMM_method)
PMM_colinfo$type <- (sapply(completeData_mice_PMM_method, class))
PMM_colinfo$na <- sapply(completeData_mice_PMM_method, function(x) sum(is.na(x)))
PMM_colinfo$notna <- sapply(completeData_mice_PMM_method, function(x) sum(!is.na(x)))
PMM_colinfo$pctdata <- PMM_colinfo$notna / (PMM_colinfo$na + PMM_colinfo$notna)
PMM_mydatasummary <- as.list(summary(completeData_mice_PMM_method))
view(PMM_colinfo)
```

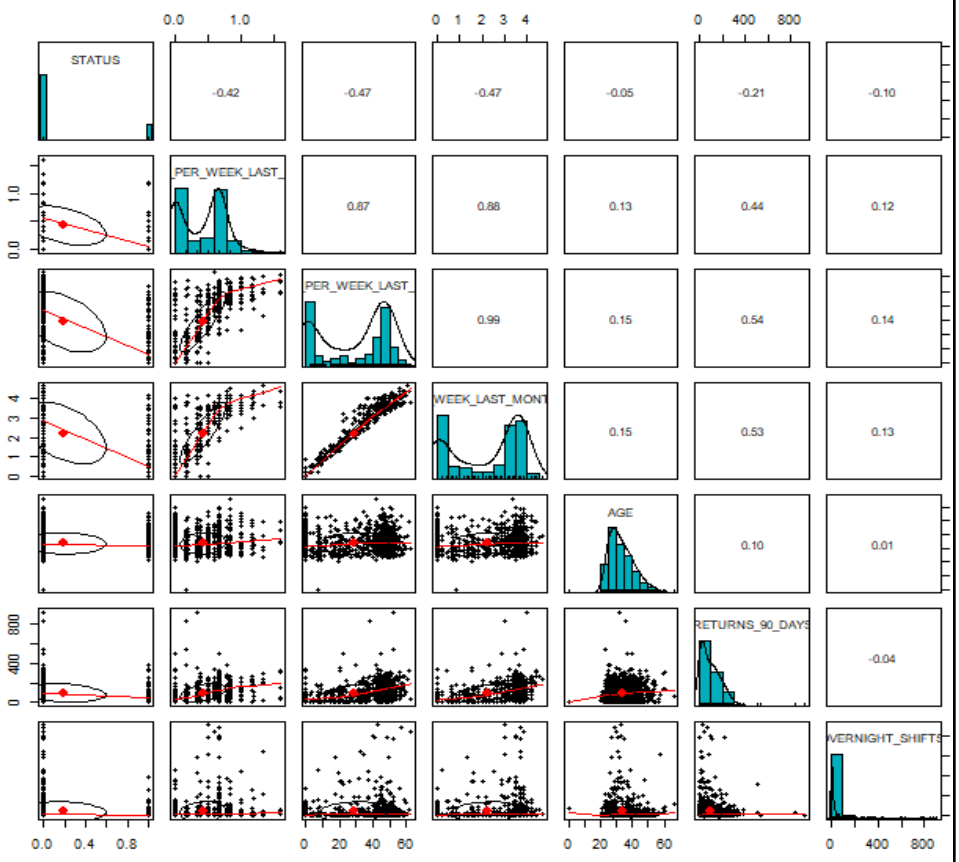
colnum	variable	type	na	notna	pctdata
1	...1	numeric	0	894	1
2	TECH	numeric	0	894	1
3	DELETE_CODE	character	0	894	1
4	AGE	numeric	0	894	1
5	STATUS	numeric	0	894	1
6	JOB_DESCRIPTION	character	0	894	1
7	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
8	CAREER_SHIFTS	numeric	0	894	1
9	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
10	WEEKEND_SHIFTS	numeric	0	894	1
11	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
12	CAREER_6_MONTH_DELIVERED_AVERAGE	numeric	0	894	1
13	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
14	CAREER_RETURNS	numeric	0	894	1
15	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
16	CAREER_DAILY_RETURN_AVERAGE	numeric	0	894	1
17	RETURNS_90_DAYS	numeric	0	894	1
18	PEICES_DELIVERED_90_DAYS	numeric	0	894	1
19	CAREER_PEICES_DELIVERED	numeric	0	894	1
20	CAREER_HOURS_WORKED	numeric	0	894	1
21	CAREER_HOURS_WORKED_DAILY_AVERAGE	numeric	0	894	1
22	SIX_MONTH_DAILY_RETURN_AVERAGE	numeric	0	894	1
23	WEEK_SHIFTS	numeric	0	894	1
24	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	numeric	0	894	1
25	CAREER_DAILY_DELIVERED_AVERAGE	numeric	0	894	1
26	OVERNIGHT_SHIFTS	numeric	0	894	1
27	DAILY_DELIVERY_PAST_MONTH_AVERAGE	numeric	0	894	1
28	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
29	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
30	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1

From top to bottom variables:

- Status.
- Weekend_shifts_per_week _last month_average.
- Hours_worked_per_week_l ast_month_average
- Shifts_per_weeek_last month_average
- Age
- Return_90_days
- Overnight_shifts.

Highly correlated variables:

- Weekend Shifts per week last month average and hours worked per week last month average (0.87).
- Weekend Shifts per week last month average and shifts per week last month average (0.88).
- Hours worked per week last month average and shifts per week last month average (0.99)




```r
# Import impute files and create correlation/scatterplot visual
completeData_mice_PMM_method <- read_excel("completeData_mice_PMM_method.xlsx")

PMM <- subset(completeData_mice_PMM_method, select = c(STATUS,
                                                       WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE,
                                                       HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE,
                                                       SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE,
                                                       AGE,
                                                       RETURNS_90_DAYS,
                                                       OVERNIGHT_SHIFTS))
pairs.panels(PMM,
             method = "pearson", # Correlation method
             hist.col = "#00AFBB",
             density = TRUE, # Show density plots
             ellipses = TRUE # Show correlation ellipses
             )
```

```
[1] FALSE
     ...1            TECH        DELETE_CODE             AGE            STATUS       JOB_DESCRIPTION    WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE CAREER_SHIFTS
 Min.   :  0.0   Min.   : 1521   Length:894         Min.   : 0.00   Min.   :0.0000   Length:894         Min.   :0.0000                             Min.   :   1.00
 1st Qu.:223.2   1st Qu.:16316   Class :character   1st Qu.:27.00   1st Qu.:0.0000   Class :character   1st Qu.:0.0000                             1st Qu.:  64.25
 Median :446.5   Median :18282   Mode  :character   Median :32.00   Median :0.0000   Mode  :character   Median :0.5000                             Median : 155.50
 Mean   :446.5   Mean   :17418                      Mean   :33.54   Mean   :0.1957                      Mean   :0.4283                             Mean   : 294.21
 3rd Qu.:669.8   3rd Qu.:19383                      3rd Qu.:38.00   3rd Qu.:0.0000                      3rd Qu.:0.6667                             3rd Qu.: 451.25
 Max.   :893.0   Max.   :20747                      Max.   :66.00   Max.   :1.0000                      Max.   :1.6000                             Max.   :1153.00
 HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE WEEKEND_SHIFTS   SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE CAREER_6_MONTH_DELIVERED_AVERAGE HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE
 Min.   : 0.00                           Min.   :  1.00   Min.   :0.000                     Min.   :   0.0                 Min.   : 0.00
 1st Qu.:10.46                           1st Qu.: 16.00   1st Qu.:1.082                     1st Qu.:  380.0                1st Qu.: 0.00
 Median :34.56                           Median : 42.00   Median :3.179                     Median :  559.3                Median :37.55
 Mean   :30.27                           Mean   : 73.67   Mean   :2.662                     Mean   :  815.5                Mean   :28.62
 3rd Qu.:46.96                           3rd Qu.:109.00   3rd Qu.:4.071                     3rd Qu.:  942.0                3rd Qu.:47.30
 Max.   :68.37                           Max.   :406.00   Max.   :5.464                     Max.   :10202.2                Max.   :63.00
 CAREER_RETURNS     SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE CAREER_DAILY_RETURN_AVERAGE RETURNS_90_DAYS  PEICES_DELIVERED_90_DAYS CAREER_PEICES_DELIVERED CAREER_HOURS_WORK
 Min.   :   1.00   Min.   :0.000                      Min.   :  1.000             Min.   :  1.00   Min.   :    10           Min.   :     0          Min.   :    0.0
 1st Qu.:  60.25   1st Qu.:0.000                      1st Qu.:  6.000             1st Qu.: 23.50   1st Qu.:  7230           1st Qu.: 10947          1st Qu.:  692.2
 Median : 220.00   Median :3.000                      Median :  9.367             Median : 80.00   Median : 22247           Median : 30976          Median : 1707.0
 Mean   : 328.36   Mean   :2.221                      Mean   :  9.853             Mean   : 98.74   Mean   : 37638           Mean   : 43975          Mean   : 3184.2
 3rd Qu.: 549.75   3rd Qu.:3.667                      3rd Qu.: 12.556             3rd Qu.:151.75   3rd Qu.: 48318           3rd Qu.: 63668          3rd Qu.: 4841.5
 Max.   :3186.00   Max.   :4.667                      Max.   :138.000             Max.   :923.00   Max.   :216042           Max.   :252142          Max.   :14153.0
 CAREER_HOURS_WORKED_DAILY_AVERAGE SIX_MONTH_DAILY_RETURN_AVERAGE  WEEK_SHIFTS    PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE CAREER_DAILY_DELIVERED_AVERAGE
 Min.   : 0.00                     Min.   :  1.00                  Min.   :1.000   Min.   :  1.00                           Min.   :    0.0
 1st Qu.:10.01                     1st Qu.:  6.10                  1st Qu.:2.000   1st Qu.:  5.00                           1st Qu.:  354.6
 Median :10.99                     Median :  8.75                  Median :2.000   Median : 16.33                           Median :  514.3
 Mean   :10.74                     Mean   :  9.42                  Mean   :1.814   Mean   : 25.68                           Mean   :  688.5
 3rd Qu.:11.71                     3rd Qu.: 11.08                  3rd Qu.:2.000   3rd Qu.: 37.00                           3rd Qu.:  766.5
 Max.   :14.46                     Max.   :271.00                  Max.   :2.000   Max.   :326.00                           Max.   :10202.2
 OVERNIGHT_SHIFTS DAILY_DELIVERY_PAST_MONTH_AVERAGE OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE
 Min.   :  1.00   Min.   :   32                    Min.   :0.0000                              Min.   :0.0000
 1st Qu.:  4.25   1st Qu.: 1690                    1st Qu.:0.0000                              1st Qu.:0.0000
 Median : 10.00   Median : 4206                    Median :0.0000                              Median :0.0000
 Mean   : 39.99   Mean   : 6136                    Mean   :0.1905                              Mean   :0.1378
 3rd Qu.: 28.75   3rd Qu.: 7961                    3rd Qu.:0.0000                              3rd Qu.:0.0000
 Max.   :900.00   Max.   :53703                    Max.   :4.3214                              Max.   :4.2000
 WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
 Min.   :0.0000
 1st Qu.:0.1786
 Median :0.6362
 Mean   :0.5620
 3rd Qu.:0.8571
 Max.   :1.8519
```

Code File	Mice_PMM_method_imputarion.R

Mean Method - Coding

- For this imputation table, we will do it column by column. Previously, we split our table because we could impute with character values on the data frame.
  Therefore, we tried using rbind and a similar method to join the data frame after imputing with mean. However, it was not possible.

```
53  colnames(df_cf)
54  summary(df_cf) # We can notice that our first six columns do not have NaN value
55  completeData_mean_imp <- df_cf
```

- This method was used for each column from seven to number 30.
  > data$x1[is.na(data$x1)] <- mean(data$x1, na.rm = **TRUE**)

```
85  # Export dataframe as a file on 11/30/2022
86  write_xlsx(completeData_mean_imp, "C:\\Users\\lewis\\Desktop\\completeData_mean_imp.xlsx")
87
```

Missing values? No, the table below has proof of this absence of NaN values.

- Source: https://statisticsglobe.com/mean-imputation-for-missing-data/

- As we can observe in the below table, we do not have missing values after performing the mean imputation.
- New imputation file was uploaded to OneDrive.

```
68 # Information about the mean dataframe
69 colnum <- (seq_len(ncol(completeData_mean_imp)))
70 Mean_imp_colinfo <- as.data.frame(colnum)
71 Mean_imp_colinfo$variable <- names(completeData_mean_imp)
72 Mean_imp_colinfo$type <- (sapply(completeData_mean_imp, class))
73 Mean_imp_colinfo$na <- sapply(completeData_mean_imp, function(x) sum(is.na(x)))
74 Mean_imp_colinfo$notna <- sapply(completeData_mean_imp, function(x) sum(!is.na(x)))
75 Mean_imp_colinfo$pctdata <- Mean_imp_colinfo$notna / (Mean_imp_colinfo$na + Mean_imp_colinfo$notna)
76 Mean_imp_mydatasummary <- as.list(summary(completeData_mean_imp))
77 view(Mean_imp_colinfo)
```

ssignment 5_v2.R* | Mean_imp_colinfo | completeData_mean_imp | PMM_colinfo | colinfo

Filter

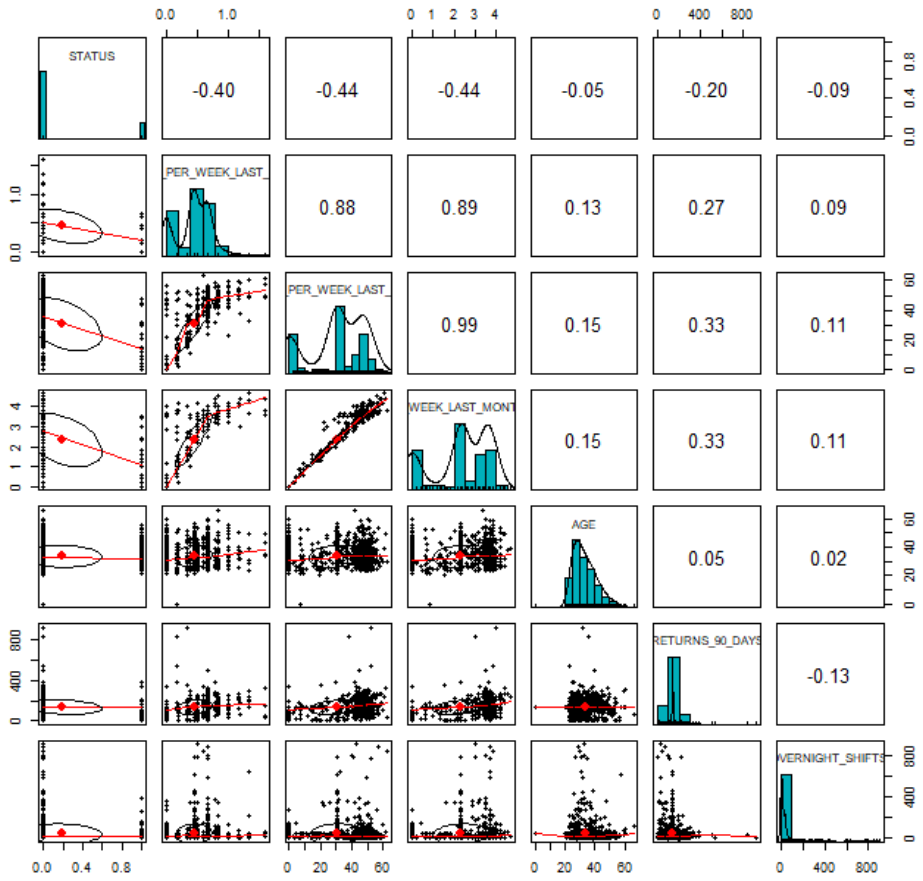
colnum	variable	type	na	notna	pctdata
1	...1	numeric	0	894	1
2	TECH	numeric	0	894	1
3	DELETE_CODE	character	0	894	1
4	AGE	numeric	0	894	1
5	STATUS	numeric	0	894	1
6	JOB_DESCRIPTION	character	0	894	1
7	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
8	CAREER_SHIFTS	numeric	0	894	1
9	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
10	WEEKEND_SHIFTS	numeric	0	894	1
11	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
12	CAREER_6_MONTH_DELIVERED_AVERAGE	numeric	0	894	1
13	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
14	CAREER_RETURNS	numeric	0	894	1
15	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
16	CAREER_DAILY_RETURN_AVERAGE	numeric	0	894	1
17	RETURNS_90_DAYS	numeric	0	894	1
18	PEICES_DELIVERED_90_DAYS	numeric	0	894	1
19	CAREER_PEICES_DELIVERED	numeric	0	894	1
20	CAREER_HOURS_WORKED	numeric	0	894	1
21	CAREER_HOURS_WORKED_DAILY_AVERAGE	numeric	0	894	1
22	SIX_MONTH_DAILY_RETURN_AVERAGE	numeric	0	894	1
23	WEEK_SHIFTS	numeric	0	894	1
24	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	numeric	0	894	1
25	CAREER_DAILY_DELIVERED_AVERAGE	numeric	0	894	1
26	OVERNIGHT_SHIFTS	numeric	0	894	1
27	DAILY_DELIVERY_PAST_MONTH_AVERAGE	numeric	0	894	1
28	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
29	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
30	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1

From top to bottom variables:

- Status.
- Weekend_shifts_per_week _last month_average.
- Hours_worked_per_week_l ast_month_average
- Shifts_per_weeek_last month_average
- Age
- Return_90_days
- Overnight_shifts.

Highly correlated variables:

- Weekend Shifts per week last month average and hours worked per week last month average (0.88).
- Weekend Shifts per week last month average and shifts per week last month average (0.89).
- Hours worked per week last month average and shifts per week last month average (0.99)




```
Mean_imp <- subset(completeData_mean_imp, select = c(STATUS,
                                                     WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE,
                                                     HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE,
                                                     SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE,
                                                     AGE,
                                                     RETURNS_90_DAYS,
                                                     OVERNIGHT_SHIFTS))
pairs.panels(Mean_imp,
             method = "pearson", # Correlation method
             hist.col = "#00AFBB",
             density = TRUE, # Show density plots
             ellipses = TRUE # Show correlation ellipses
)
```

```
completeData_mean_imp$"WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE"[is.na(completeData_mean_imp$"WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE")] <- mean(completeData_mean_imp$"WEEKEND_SHIFTS_PER_WEEK
completeData_mean_imp$"CAREER_SHIFTS"[is.na(completeData_mean_imp$"CAREER_SHIFTS")] <- mean(completeData_mean_imp$"CAREER_SHIFTS", na.rm = TRUE)
completeData_mean_imp$"HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE"[is.na(completeData_mean_imp$"HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE")] <- mean(completeData_mean_imp$"HOURS_WORKED_PER_WEEK_SIX_MON
completeData_mean_imp$"WEEKEND_SHIFTS"[is.na(completeData_mean_imp$"WEEKEND_SHIFTS")] <- mean(completeData_mean_imp$"WEEKEND_SHIFTS", na.rm = TRUE)
completeData_mean_imp$"SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE"[is.na(completeData_mean_imp$"SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE")] <- mean(completeData_mean_imp$"SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE", na.rm
```

```
     ...1            TECH       DELETE_CODE            AGE            STATUS       JOB_DESCRIPTION    WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE CAREER_SHIFTS
Min.   :  0.0   Min.   : 1521   Length:894         Min.   : 0.00   Min.   :0.0000   Length:894         Min.   :0.0000                             Min.   :   1.00
1st Qu.:223.2   1st Qu.:16316   Class :character   1st Qu.:27.00   1st Qu.:0.0000   Class :character   1st Qu.:0.3333                             1st Qu.:  71.25
Median :446.5   Median :18282   Mode  :character   Median :32.00   Median :0.0000   Mode  :character   Median :0.4490                             Median : 189.50
Mean   :446.5   Mean   :17418                      Mean   :33.54   Mean   :0.1957                      Mean   :0.4490                             Mean   : 305.38
3rd Qu.:669.8   3rd Qu.:19383                      3rd Qu.:38.00   3rd Qu.:0.0000                      3rd Qu.:0.6667                             3rd Qu.: 440.00
Max.   :893.0   Max.   :20747                      Max.   :66.00   Max.   :1.0000                      Max.   :1.6000                             Max.   :1153.00
HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE WEEKEND_SHIFTS   SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE CAREER_6_MONTH_DELIVERED_AVERAGE HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE
Min.   : 0.00                           Min.   :  1.00   Min.   :0.000                     Min.   :    0.0                 Min.   : 0.00
1st Qu.:25.76                           1st Qu.: 20.00   1st Qu.:2.222                     1st Qu.:  458.1                 1st Qu.:23.45
Median :32.08                           Median : 54.00   Median :2.758                     Median :  668.0                 Median :30.38
Mean   :32.08                           Mean   : 77.73   Mean   :2.758                     Mean   :  668.0                 Mean   :30.38
3rd Qu.:44.98                           3rd Qu.:105.75   3rd Qu.:3.857                     3rd Qu.:  668.0                 3rd Qu.:46.17
Max.   :68.37                           Max.   :406.00   Max.   :5.464                     Max.   :10202.2                 Max.   :63.00
CAREER_RETURNS   SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE CAREER_DAILY_RETURN_AVERAGE RETURNS_90_DAYS PEICES_DELIVERED_90_DAYS CAREER_PEICES_DELIVERED CAREER_HOURS_WORKED
Min.   :   1.0   Min.   :0.000                     Min.   :  1.000              Min.   :  1.0   Min.   :    10          Min.   :     0          Min.   :    0.0
1st Qu.:  88.5   1st Qu.:1.875                    1st Qu.:  6.839              1st Qu.:124.0   1st Qu.: 21104          1st Qu.: 22184          1st Qu.:  728.2
Median : 321.0   Median :2.332                    Median : 10.126              Median :137.9   Median : 34703          Median : 46233          Median : 1890.0
Mean   : 349.1   Mean   :2.332                    Mean   : 10.126              Mean   :137.9   Mean   : 34703          Mean   : 46233          Mean   : 3251.3
3rd Qu.: 541.0   3rd Qu.:3.500                    3rd Qu.: 12.445              3rd Qu.:140.8   3rd Qu.: 34703          3rd Qu.: 48666          3rd Qu.: 4725.8
Max.   :3186.0   Max.   :4.667                    Max.   :138.000              Max.   :923.0   Max.   :216042          Max.   :252142          Max.   :14153.0
CAREER_HOURS_WORKED_DAILY_AVERAGE SIX_MONTH_DAILY_RETURN_AVERAGE  WEEK_SHIFTS    PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE CAREER_DAILY_DELIVERED_AVERAGE OVERNIGHT_SHIFTS
Min.   : 0.00                   Min.   :  1.000                  Min.   :1.000   Min.   :  1.00                           Min.   :    0.0                 Min.   :  1.00
1st Qu.:10.12                   1st Qu.:  8.755                  1st Qu.:1.758   1st Qu.: 35.00                           1st Qu.:  429.0                 1st Qu.:  5.00
Median :10.93                   Median : 10.385                  Median :1.758   Median : 39.96                           Median :  635.6                 Median : 13.00
Mean   :10.76                   Mean   : 10.385                  Mean   :1.758   Mean   : 39.96                           Mean   :  635.6                 Mean   : 41.24
3rd Qu.:11.70                   3rd Qu.: 10.385                  3rd Qu.:1.758   3rd Qu.: 39.96                           3rd Qu.:  635.6                 3rd Qu.: 41.24
Max.   :14.46                   Max.   :271.000                  Max.   :2.000   Max.   :326.00                           Max.   :10202.2                 Max.   :900.00
DAILY_DELIVERY_PAST_MONTH_AVERAGE OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE
Min.   :   32                 Min.   :0.0000                                 Min.   :0.0000                              Min.   :0.0000
1st Qu.: 6440                 1st Qu.:0.0000                                 1st Qu.:0.0000                              1st Qu.:0.4286
Median : 8220                 Median :0.0000                                 Median :0.0000                              Median :0.5804
Mean   : 8220                 Mean   :0.1372                                 Mean   :0.1055                              Mean   :0.5804
3rd Qu.: 8220                 3rd Qu.:0.1372                                 3rd Qu.:0.1055                              3rd Qu.:0.7857
Max.   :53703                 Max.   :4.3214                                 Max.   :4.2000                              Max.   :1.8519
```

Mean imputation file:	Mean_imputation_v2.R

Median interpolation method : We chose to perform this method because the client requested it in order to deal with issues that arise from not having a normal distribution. This method replaces the missing value with the median value of the entire feature column.	

Nulls Before

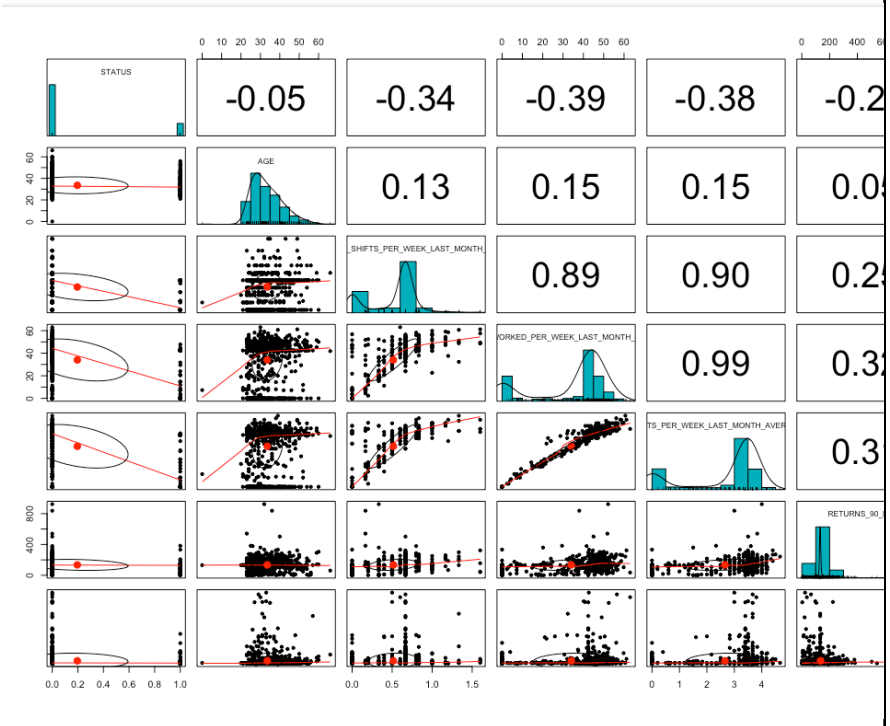
	colnum	variable	type	na
1	1	...1	numeric	0
2	2	TECH	numeric	0
3	3	DELETE_CODE	character	0
4	4	AGE	numeric	0
5	5	STATUS	numeric	0
6	6	JOB_DESCRIPTION	character	0
7	7	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	273
8	8	CAREER_SHIFTS	numeric	59
9	9	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	numeric	273
10	10	WEEKEND_SHIFTS	numeric	79
11	11	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	273
12	12	CAREER_6_MONTH_DELIVERED_AVERAGE	numeric	339
13	13	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	numeric	273
14	14	CAREER_RETURNS	numeric	97
15	15	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	273
16	16	CAREER_DAILY_RETURN_AVERAGE	numeric	97
17	17	RETURNS_90_DAYS	numeric	402
18	18	PEICES_DELIVERED_90_DAYS	numeric	404
19	19	CAREER_PEICES_DELIVERED	numeric	304
20	20	CAREER_HOURS_WORKED	numeric	39
21	21	CAREER_HOURS_WORKED_DAILY_AVERAGE	numeric	39
22	22	SIX_MONTH_DAILY_RETURN_AVERAGE	numeric	337
23	23	WEEK_SHIFTS	numeric	687
24	24	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	numeric	470
25	25	CAREER_DAILY_DELIVERED_AVERAGE	numeric	301
26	26	OVERNIGHT_SHIFTS	numeric	110
27	27	DAILY_DELIVERY_PAST_MONTH_AVERAGE	numeric	476
28	28	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	273

Code

```r
#load necessary packages
library(dplyr)
library(tidyr)
library(writexl)
library(psych)
library(ggplot2)
library(lmtest)
library(nortest)
library(tidyverse)
library(car)

#load in data
cfdata <- ops_turnover_feature_set

#remove non-numeric columns
cfdata1 <- cfdata[,-c(3,3)]
cfdata2 <- cfdata1[,-c(5,5)]

#find missing values based off median
cfdata3 <- cfdata2 %>% mutate(across(where(is.numeric), ~replace_na(., median(., na.rm=TRUE))))

#export as excel file
write_xlsx(cfdata3,"Desktop/Analytics Practicum/ops_turnover_feature_set.xlsx")

#find correlation
aadd<-select(cfdata3 ,STATUS ,AGE,WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE,
             HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE, SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE, RETU
             OVERNIGHT_SHIFTS)

pairs.panels(aadd,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
)
```

Nulls After

	colnum	variable	type	na
1	1	...1	numeric	0
2	2	TECH	numeric	0
3	3	AGE	numeric	0
4	4	STATUS	numeric	0
5	5	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0
6	6	CAREER_SHIFTS	numeric	0
7	7	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0
8	8	WEEKEND_SHIFTS	numeric	0
9	9	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0
10	10	CAREER_6_MONTH_DELIVERED_AVERAGE	numeric	0
11	11	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0
12	12	CAREER_RETURNS	numeric	0
13	13	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0
14	14	CAREER_DAILY_RETURN_AVERAGE	numeric	0
15	15	RETURNS_90_DAYS	numeric	0
16	16	PEICES_DELIVERED_90_DAYS	numeric	0
17	17	CAREER_PEICES_DELIVERED	numeric	0
18	18	CAREER_HOURS_WORKED	numeric	0
19	19	CAREER_HOURS_WORKED_DAILY_AVERAGE	numeric	0
20	20	SIX_MONTH_DAILY_RETURN_AVERAGE	numeric	0
21	21	WEEK_SHIFTS	numeric	0
22	22	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	numeric	0
23	23	CAREER_DAILY_DELIVERED_AVERAGE	numeric	0
24	24	OVERNIGHT_SHIFTS	numeric	0
25	25	DAILY_DELIVERY_PAST_MONTH_AVERAGE	numeric	0
26	26	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0
27	27	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0
28	28	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0

Correlation: The following variables show correlation with the status variable: • Weekend_Shifts_Per_Week_Last_Month_Average • Hours_Worked_Per_Week_Last_Month_Average • Shifts_Per_Week_Last_Month_Average	
Code File	median_imputationv2.R

Piecewise Polynomial Interpolation Method	
Why PieceWise	PieceWise can be used when the interval between [a,b] is too large. To trough this method the interval is broken down into “subintervals”. Since we have a lot of missing values this method will help since we will have very large intervals between the data used to predict the NA values.

Nulls before

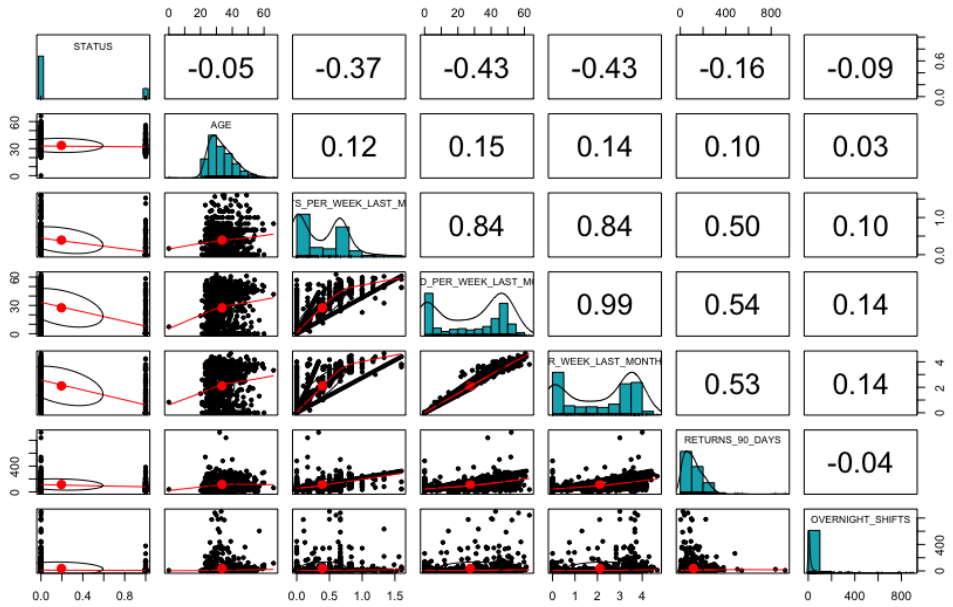
	FEATURE	NULL_COUNT	NULL_RATIO
0	Unnamed: 0	0	0.000000
1	TECH	0	0.000000
2	DELETE_CODE	0	0.000000
3	AGE	0	0.000000
4	STATUS	0	0.000000
5	JOB_DESCRIPTION	0	0.000000
6	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	273	0.305369
7	CAREER_SHIFTS	59	0.065996
8	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	273	0.305369
9	WEEKEND_SHIFTS	79	0.088367
10	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	273	0.305369
11	CAREER_6_MONTH_DELIVERED_AVERAGE	339	0.379195
12	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	273	0.305369
13	CAREER_RETURNS	97	0.108501
14	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	273	0.305369
15	CAREER_DAILY_RETURN_AVERAGE	97	0.108501
16	RETURNS_90_DAYS	402	0.449664
17	PEICES_DELIVERED_90_DAYS	404	0.451902
18	CAREER_PEICES_DELIVERED	304	0.340045
19	CAREER_HOURS_WORKED	39	0.043624
20	CAREER_HOURS_WORKED_DAILY_AVERAGE	39	0.043624
21	SIX_MONTH_DAILY_RETURN_AVERAGE	337	0.376957
22	WEEK_SHIFTS	687	0.768456
23	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	470	0.525727
24	CAREER_DAILY_DELIVERED_AVERAGE	301	0.336689
25	OVERNIGHT_SHIFTS	110	0.123043
26	DAILY_DELIVERY_PAST_MONTH_AVERAGE	476	0.532438
27	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	273	0.305369
28	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	273	0.305369
29	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	273	0.305369

Interpolation code

```
#interpolate dataset

interpolated_ops_features = ops_features.interpolate(method = 'piecewise_polynomial')
inter_null_df = pd.DataFrame(interpolated_ops_features.isnull().sum())
inter_null_df.reset_index(level = 0, inplace = True)
inter_null_df.rename(columns = {0: 'NULL_COUNT', 'index': 'FEATURE'}, inplace = True)
inter_null_df['NULL_RATIO'] = inter_null_df['NULL_COUNT']/interpolated_ops_features.shape[0]
```

Nulls after

	FEATURE	NULL_COUNT	NULL_RATIO
0	Unnamed: 0	0	0.000000
1	TECH	0	0.000000
2	DELETE_CODE	0	0.000000
3	AGE	0	0.000000
4	STATUS	0	0.000000
5	JOB_DESCRIPTION	0	0.000000
6	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	1	0.001119
7	CAREER_SHIFTS	0	0.000000
8	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	1	0.001119
9	WEEKEND_SHIFTS	1	0.001119
10	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	1	0.001119
11	CAREER_6_MONTH_DELIVERED_AVERAGE	0	0.000000
12	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	1	0.001119
13	CAREER_RETURNS	0	0.000000
14	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	1	0.001119
15	CAREER_DAILY_RETURN_AVERAGE	0	0.000000
16	RETURNS_90_DAYS	0	0.000000
17	PEICES_DELIVERED_90_DAYS	0	0.000000
18	CAREER_PEICES_DELIVERED	0	0.000000
19	CAREER_HOURS_WORKED	0	0.000000
20	CAREER_HOURS_WORKED_DAILY_AVERAGE	0	0.000000
21	SIX_MONTH_DAILY_RETURN_AVERAGE	0	0.000000
22	WEEK_SHIFTS	8	0.008949
23	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	0	0.000000
24	CAREER_DAILY_DELIVERED_AVERAGE	0	0.000000
25	OVERNIGHT_SHIFTS	0	0.000000
26	DAILY_DELIVERY_PAST_MONTH_AVERAGE	0	0.000000
27	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	1	0.001119
28	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	1	0.001119
29	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	1	0.001119

Correlation We can see there is a correlation between status and • Weekend_Shifts_Per_Week_Last_Month_Average • Hours_Worked_Per_Week_Last_Month_Average • Shifts_Per_Week_Last_Month_Average That might be worth looking into as we build a model to predict the status variable for the employees.	
Code File	Piecewise Interpolation.ipynb PieceWise Correlation.R

Random Forest Imputation Method

Before Imputation

	colnum	variable	type	na	notna	pctdata
1	1	X	integer	0	894	1.0000000
2	2	TECH	numeric	0	894	1.0000000
3	3	DELETE_CODE	character	0	894	1.0000000
4	4	AGE	integer	0	894	1.0000000
5	5	STATUS	integer	0	894	1.0000000
6	6	JOB_DESCRIPTION	character	0	894	1.0000000
7	7	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	273	621	0.6946309
8	8	CAREER_SHIFTS	numeric	59	835	0.9340045
9	9	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	numeric	273	621	0.6946309
10	10	WEEKEND_SHIFTS	numeric	79	815	0.9116331
11	11	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	273	621	0.6946309
12	12	CAREER_6_MONTH_DELIVERED_AVERAGE	numeric	339	555	0.6208054
13	13	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	numeric	273	621	0.6946309
14	14	CAREER_RETURNS	numeric	97	797	0.8914989
15	15	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	273	621	0.6946309
16	16	CAREER_DAILY_RETURN_AVERAGE	numeric	97	797	0.8914989
17	17	RETURNS_90_DAYS	numeric	402	492	0.5503356
18	18	PEICES_DELIVERED_90_DAYS	numeric	404	490	0.5480984
19	19	CAREER_PEICES_DELIVERED	numeric	304	590	0.6599553
20	20	CAREER_HOURS_WORKED	numeric	39	855	0.9563758
21	21	CAREER_HOURS_WORKED_DAILY_AVERAGE	numeric	39	855	0.9563758
22	22	SIX_MONTH_DAILY_RETURN_AVERAGE	numeric	337	557	0.6230425
23	23	WEEK_SHIFTS	numeric	687	207	0.2315436
24	24	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	numeric	470	424	0.4742729
25	25	CAREER_DAILY_DELIVERED_AVERAGE	numeric	301	593	0.6633110
26	26	OVERNIGHT_SHIFTS	numeric	110	784	0.8769575
27	27	DAILY_DELIVERY_PAST_MONTH_AVERAGE	numeric	476	418	0.4675615
28	28	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	273	621	0.6946309
29	29	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	273	621	0.6946309
30	30	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	273	621	0.6946309

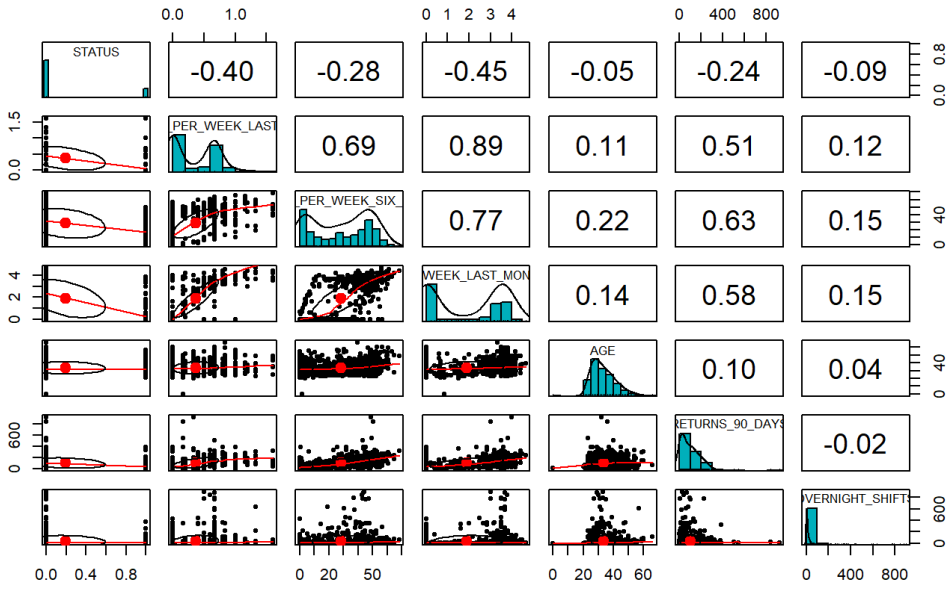
Code

```
library(mice)
set.seed(67)
packageVersion('mice')#v3.15.0
miceMod <- mice(datacity[,!names(datacity)%in%"x"],method='rf')
miceoutput <- complete(miceMod)
summary(miceoutput)
```

After Imputation	

	colnum2	variable	type	na	notna	pctdata
1	1	X	integer	0	894	1
2	2	TECH	numeric	0	894	1
3	3	DELETE_CODE	character	0	894	1
4	4	AGE	integer	0	894	1
5	5	STATUS	integer	0	894	1
6	6	JOB_DESCRIPTION	character	0	894	1
7	7	WEEKEND_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
8	8	CAREER_SHIFTS	numeric	0	894	1
9	9	HOURS_WORKED_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
10	10	WEEKEND_SHIFTS	numeric	0	894	1
11	11	SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
12	12	CAREER_6_MONTH_DELIVERED_AVERAGE	numeric	0	894	1
13	13	HOURS_WORKED_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
14	14	CAREER_RETURNS	numeric	0	894	1
15	15	SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
16	16	CAREER_DAILY_RETURN_AVERAGE	numeric	0	894	1
17	17	RETURNS_90_DAYS	numeric	0	894	1
18	18	PEICES_DELIVERED_90_DAYS	numeric	0	894	1
19	19	CAREER_PEICES_DELIVERED	numeric	0	894	1
20	20	CAREER_HOURS_WORKED	numeric	0	894	1
21	21	CAREER_HOURS_WORKED_DAILY_AVERAGE	numeric	0	894	1
22	22	SIX_MONTH_DAILY_RETURN_AVERAGE	numeric	0	894	1
23	23	WEEK_SHIFTS	numeric	0	894	1
24	24	PAST_MONTH_PIECES_RETURNED_DAILY_AVERAGE	numeric	0	894	1
25	25	CAREER_DAILY_DELIVERED_AVERAGE	numeric	0	894	1
26	26	OVERNIGHT_SHIFTS	numeric	0	894	1
27	27	DAILY_DELIVERY_PAST_MONTH_AVERAGE	numeric	0	894	1
28	28	OVERNIGHT_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1
29	29	OVERNIGHT_SHIFTS_PER_WEEK_LAST_MONTH_AVERAGE	numeric	0	894	1
30	30	WEEKEND_SHIFTS_PER_WEEK_SIX_MONTH_AVERAGE	numeric	0	894	1

Random Forest explanation	Random forest imputation is a machine learning technique which can accommodate nonlinearities and interactions and does not require a particular regression model to be specified.
Something can be improved	Can use set seeds to find the best imputation by random forest.
Code File	random forest imputation V2.r

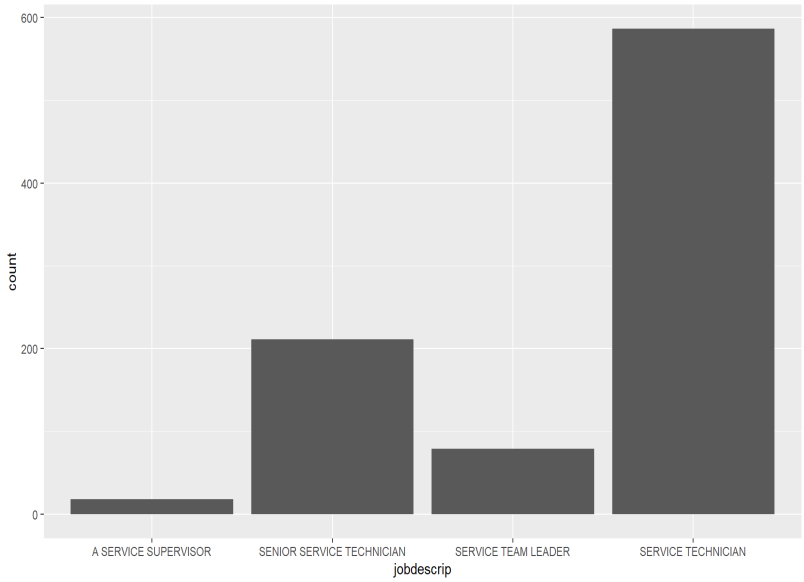
Correlation Strong correlation between: • Weekend_Shifts_Per_Week_Last_Month_Average • Shifts_Per_Week_Last_Month_Average	

We will be utilizing these 5 interpolated datasets over the original as these will give us a greater chance at building a more complete and effective model.

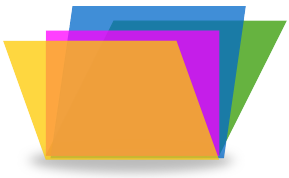
Data Munging

At this moment, the student team has not added new variables from external sources or created from the original data received. However, we have changed the datatype of the variable “job description” from a character to a factor. We are also planning to add new variables after receiving the datasets from the City Furniture employees’ survey.

See the data type of each variable and create a new column with name “job descrip”, with data type of factor.	``` datacitysummary <- as.list(summary(datacity)) datacity$jobdescrip <- as.factor(datacity$JOB_DESCRIPTION) class(datacity$type) ```
Using jobdescrib to create a plot	``` datacity$jobdescrib <- factor(datacity$jobdescrip,levels = c(1,2,3,4), labels = c("A SERVICE SUPERVISOR", "SERVICE TEAM LEADER", "SERVICE TECHNICIAN", "SENIOR SERVICE TECHNICIAN")) ggplot(datacity,aes(jobdescrip))+ geom_bar() ```

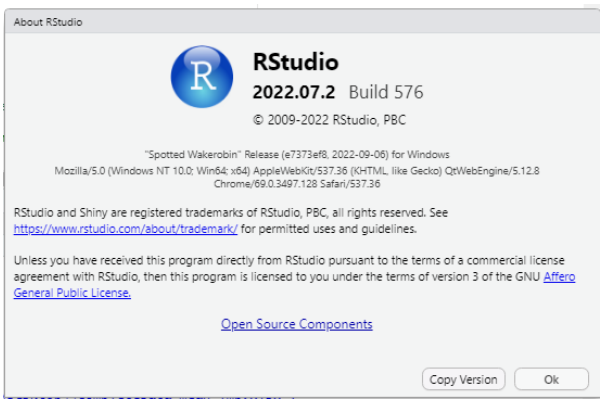
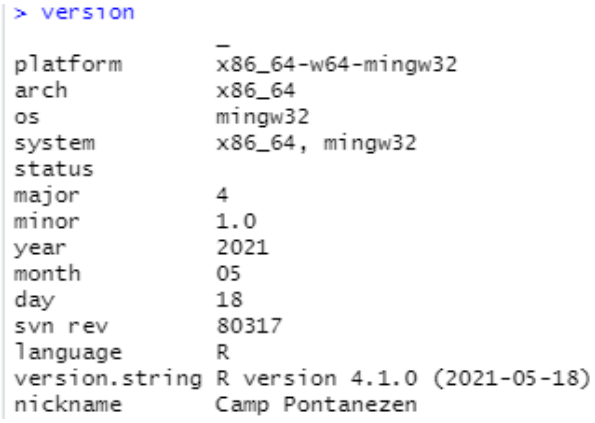
Service Supervisor with 18 employees (2.01%). Senior Service Technician with 211 employees (23.60%). Service Team Leader with 79 employees (8.84%). Service Technicians with 586 employees (65.54%).	

Development Workflow for Modeling Dataset

Import Initial Dataset	**Data Understanding**	**Data Preparation**	**Develop Final Tables That Will be Used in Modeling**
			
Step 1	**Step 2**	**Step 3**	**Step 4**

Import Excel file from sponsor's email on November 11th, 2022: ops_turnover_feature_set.csv into the shared and secure OneDrive with the sponsor.The uploaded as a dataframe in RStudio.	During the whole month of November, the student team performed the following actions: describe and explore data for assignments 1,2, and 3 in RStudio. Mice, Psych, Hmisc, Writexl, and VIM packages were installed in RStudio to perform visualization for the missing values.	Imputation methods were used to fill missing values: Mean, Median, Mice (method: Predictive mean matching), Piecewise, Random Forest. Folder with imputation methods codes: Interpolation Methods Code	The student team is waiting for a new dataset from employees' surveys to analyze. After performing imputation methods, the files of the interpolated datasets were created and uploaded to the shared OneDrive.

Documenting Data Prep Workflow and Code

Lewis Rincon Castano - code	
	

```
> sessionInfo()
R version 4.1.0 (2021-05-18)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22621)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252  LC_CTYPE=English_United States.1252    LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C                           LC_TIME=English_United States.1252

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] psych_2.2.9      writexl_1.4.1    VIM_6.2.2        colorspace_2.0-3 forcats_0.5.2    stringr_1.4.1    dplyr_1.0.10
 [8] purrr_0.3.5      readr_2.1.3      tidyr_1.2.1      tibble_3.1.8     tidyverse_1.3.2  Hmisc_4.7-1      ggplot2_3.4.0
[15] Formula_1.2-4    survival_3.4-0   lattice_0.20-44  mice_3.15.0      readxl_1.4.1

loaded via a namespace (and not attached):
 [1] nlme_3.1-152        fs_1.5.2            lubridate_1.9.0     RColorBrewer_1.1-3  httr_1.4.4          tools_4.1.0
 [7] backports_1.4.1     utf8_1.2.2          R6_2.5.1            rpart_4.1.19        DBI_1.1.3           nnet_7.3-16
[13] withr_2.5.0         sp_1.5-1            mnormt_2.1.1        tidyselect_1.2.0    gridExtra_2.3       compiler_4.1.0
[19] cli_3.4.1           rvest_1.0.3         htmlTable_2.4.1     xml2_1.3.3          scales_1.2.1        checkmate_2.1.0
[25] lmtest_0.9-40       DEoptimR_1.0-11     robustbase_0.95-0   proxy_0.4-27        digest_0.6.30       foreign_0.8-81
[31] base64enc_0.1-3     jpeg_0.1-9          pkgconfig_2.0.3     htmltools_0.5.3     dbplyr_2.2.1        fastmap_1.1.0
[37] htmlwidgets_1.5.4   rlang_1.0.6         rstudioapi_0.14     generics_0.1.3      zoo_1.8-11          jsonlite_1.8.3
[43] car_3.1-1           googlesheets4_1.0.1 magrittr_2.0.3      interp_1.1-3        Matrix_1.5-3        Rcpp_1.0.9
[49] munsell_0.5.0       fansi_1.0.3         abind_1.4-5         lifecycle_1.0.3     stringi_1.7.8       carData_3.0-5
[55] MASS_7.3-58.1       parallel_4.1.0      crayon_1.5.2        deldir_1.0-6        haven_2.5.1         splines_4.1.0
[61] hms_1.1.2           knitr_1.40          pillar_1.8.1        ranger_0.14.1       boot_1.3-28         reprex_2.0.2
[67] glue_1.6.2          latticeExtra_0.6-30 laeken_0.5.2        data.table_1.14.4   modelr_0.1.10       vcd_1.4-10
[73] png_0.1-7           vctrs_0.5.0         tzdb_0.3.0          cellranger_1.1.0    gtable_0.3.1        assertthat_0.2.1
[79] xfun_0.34           broom_1.0.1         e1071_1.7-12        class_7.3-19        googledrive_2.0.0   gargle_1.2.1
[85] cluster_2.1.2       timechange_0.1.1    ellipsis_0.3.2
```

ADD CODE LINES?