CMPE 343 Logistic Regression Analysis

I chose the HR analysis dataset. This, examines the factors influencing employee turnover in an organization using the HR Analysis dataset

General information about the dataset.

```
library(ggplot2)

# Veri kümesini yükle
data <- read.csv("/Users/sudenurerturk/Downloads/HR_comma_sep.csv")

# İlk birkaç satırı kontrol et
head(data)

# Veri yapısını kontrol et
str(data)

# Özet istatistikler
summary(data)

# Eksik verileri kontrol etme
colSums(is.na(data))
```

```
> head(data)
  satisfaction_level last_evaluation number_project average_montly_hours
1               0.38            0.53              2                   157
2               0.80            0.86              5                   262
3               0.11            0.88              7                   272
4               0.72            0.87              5                   223
5               0.37            0.52              2                   159
6               0.41            0.50              2                   153
  time_spend_company Work_accident left promotion_last_5years Department salary
1                  3             0    1                     0      sales    low
2                  6             0    1                     0      sales medium
3                  4             0    1                     0      sales medium
4                  5             0    1                     0      sales    low
5                  3             0    1                     0      sales    low
6                  3             0    1                     0      sales    low
>
```

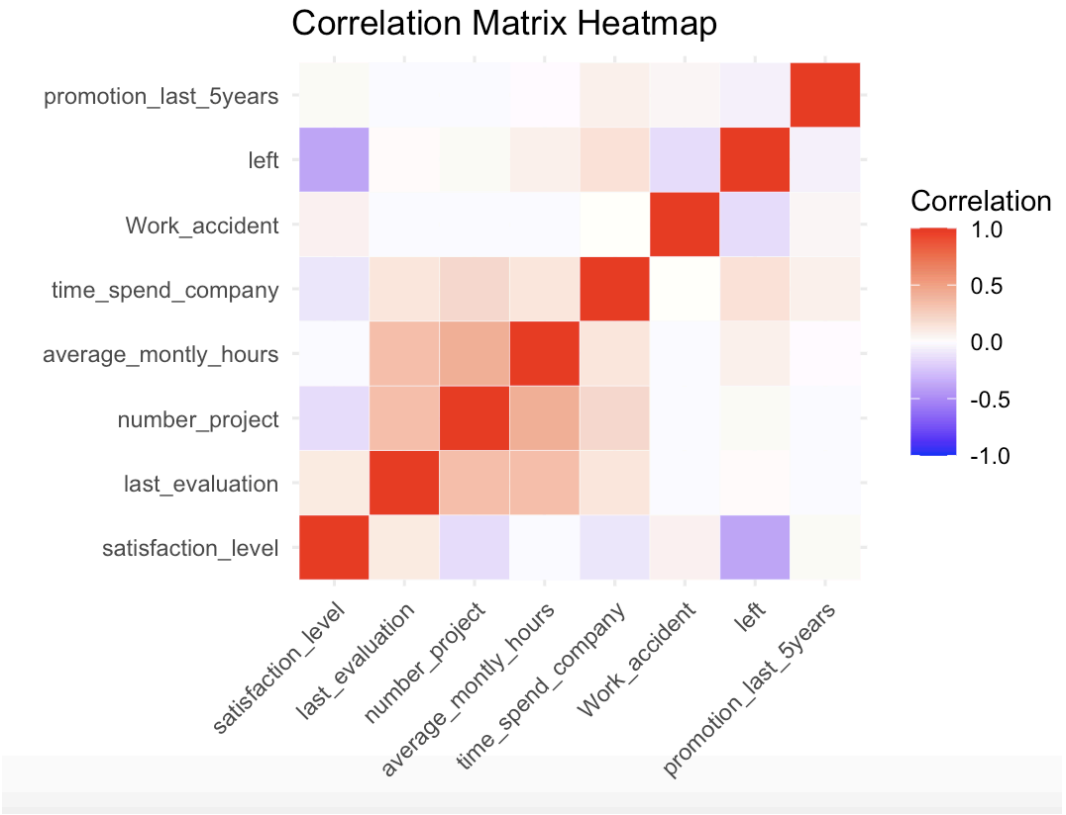
```
> # Özet istatistikler
> summary(data)
 satisfaction_level last_evaluation  number_project  average_montly_hours
 Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0
 1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
 Median :0.6400     Median :0.7200   Median :4.000   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
 time_spend_company Work_accident         left        promotion_last_5years
 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



>
> # Eksik verileri kontrol etme
> colSums(is.na(data))
   satisfaction_level       last_evaluation        number_project
                    0                     0                     0
 average_montly_hours    time_spend_company         Work_accident
                    0                     0                     0
                 left promotion_last_5years            Department
                    0                     0                     0
               salary
                    0
```

Correlation Analysis:

- **Satisfaction Level**: Strong negative correlation with turnover, indicating that dissatisfied employees are more likely to leave.
- **Time Spent at the Company**: Positive correlation, suggesting longer-tenured employees may experience burnout or stagnation.
- **Number of Projects**: Moderate correlation with turnover, highlighting workload as a potential factor.



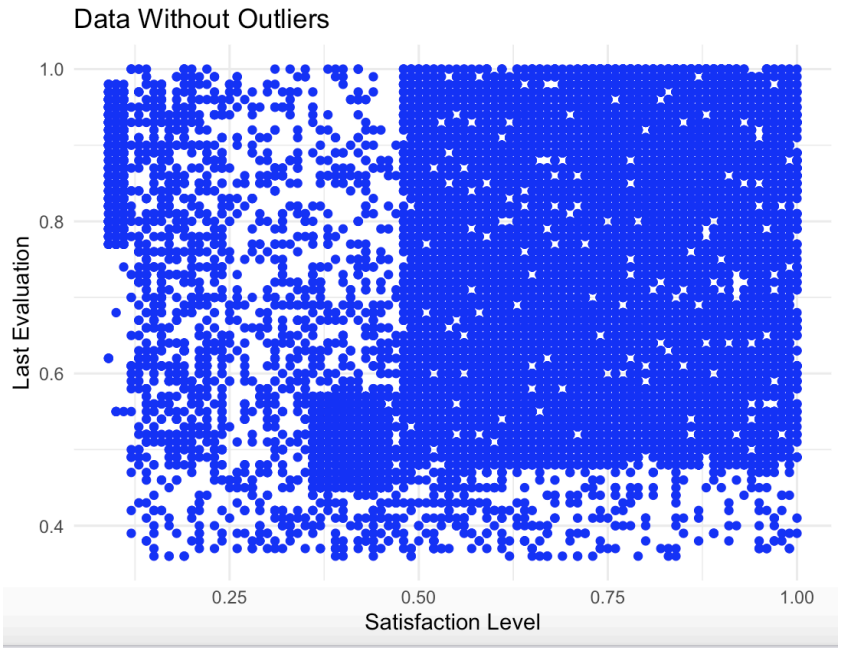
Confusion Matrix

```
> # Karışıklık matrisi (Confusion Matrix)
> confusion_matrix <- table(TrueClass = y_test, PredClass = predicted_class)
> print(confusion_matrix)
         PredClass
TrueClass    0    1
        0 1985  187
        1  398  296
>
```

Outlier Detection

Outliers were identified using Z-scores for numerical variables (e.g., satisfaction level, last evaluation, average monthly hours):

- Employees with extreme **average monthly hours** or unusually high/low **satisfaction levels** were flagged as outliers.
- Most outliers belonged to the high-risk turnover group, suggesting the need for workload adjustment or targeted retention strategies.

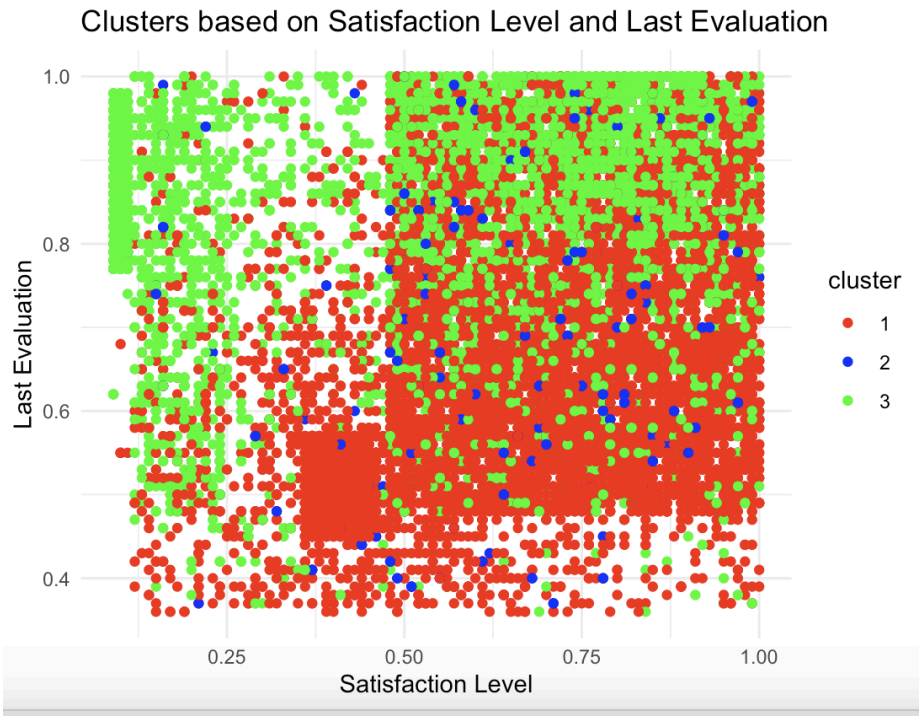


Clustring

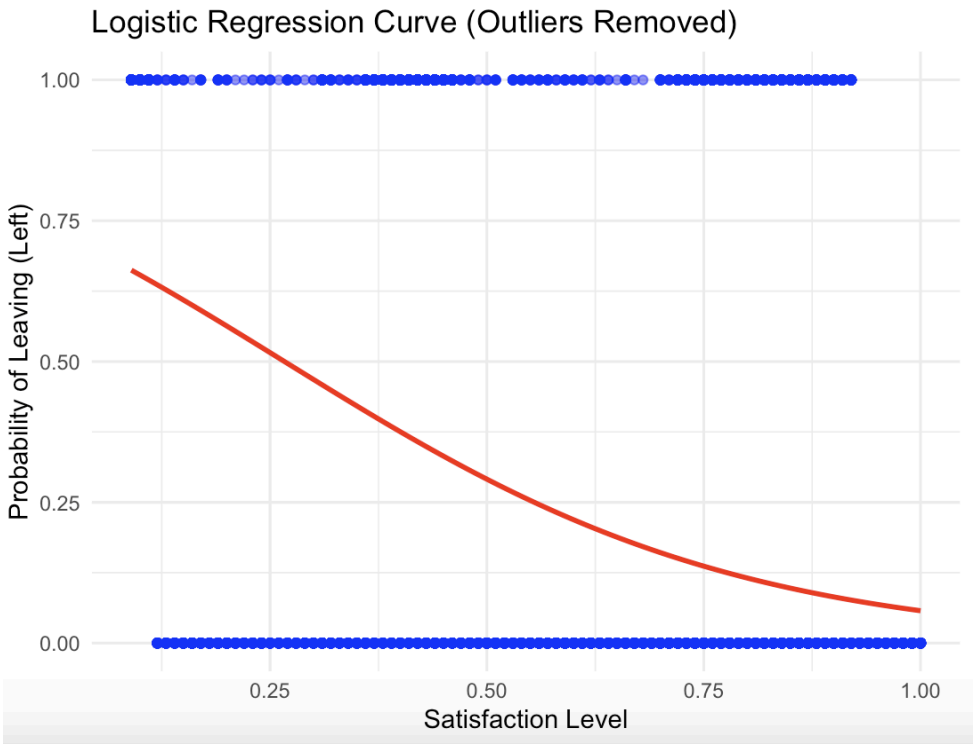
Clusters Based on Satisfaction Level and Last Evaluation

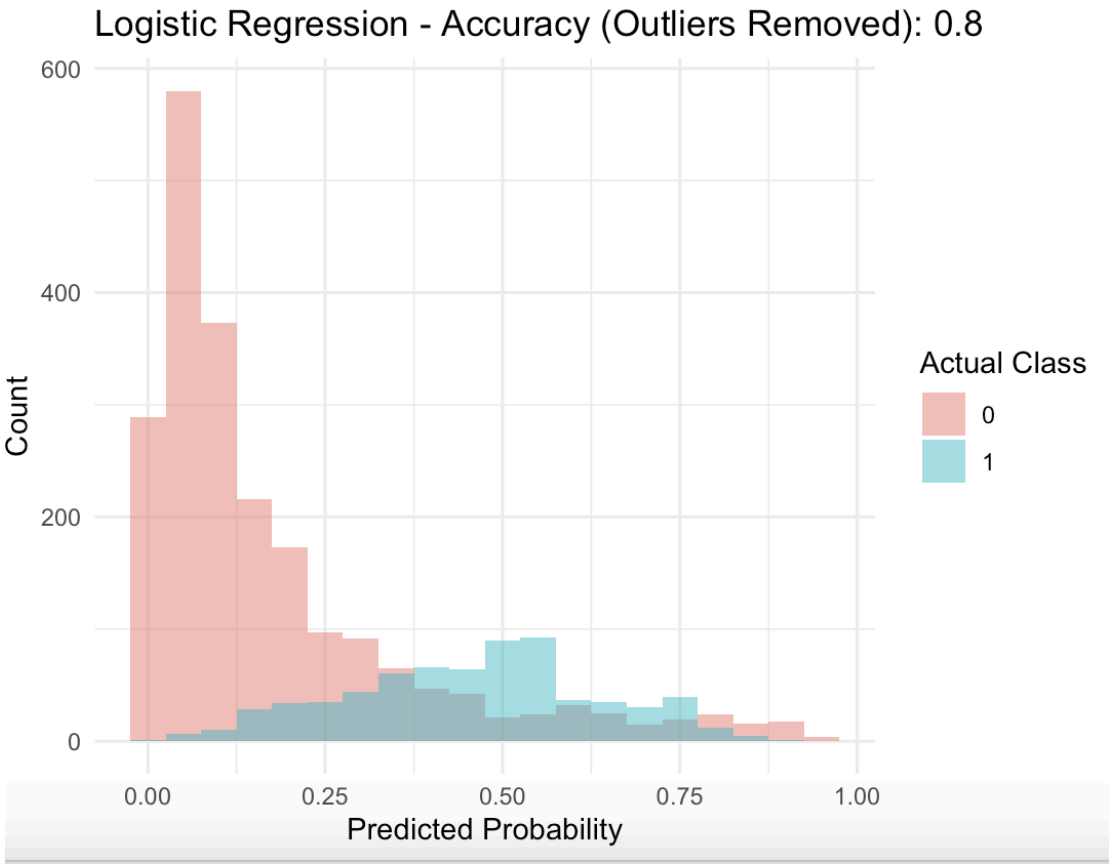
Using K-means clustering (with 3 clusters), employees were grouped based on their satisfaction levels and last evaluation scores:

- **Cluster 1**: High satisfaction, high evaluations (most likely to stay).
- **Cluster 2**: Low satisfaction, low evaluations (highest turnover risk).
- **Cluster 3**: Mixed satisfaction and evaluation levels (moderate risk). These clusters highlight distinct behavioral patterns among employees, aiding targeted interventions.



Logistic Regression Model





Accuracy: The model achieved an accuracy of **0.8** (80%), indicating a reasonably good performance in predicting employee turnover.
Distribution of Predicted Probabilities:

- The **pink histogram (Actual Class = 0)** represents employees who stayed. Most probabilities for this group are concentrated towards the lower range (0.0–0.3), showing that the model correctly predicts them as likely to stay.
- The **blue histogram (Actual Class = 1)** represents employees who left. These probabilities are more spread out but peak around the higher probability range (0.6–1.0), showing the model identifies some leavers well.

Overlap: There is some overlap between the two classes in the range of **0.3–0.6**, indicating ambiguity in prediction. This could lead to false positives (predicting an employee will leave when they stay) or false negatives (predicting an employee will stay when they leave).
Insights:

- Employees with predicted probabilities closer to 0 are predominantly stayers (Class 0).
- Employees with predicted probabilities closer to 1 are predominantly leavers (Class 1).
- The model's ability to classify is more accurate at the extremes (near 0 or 1) but less certain in the middle range.

Sude Nur Ertürk 120200039