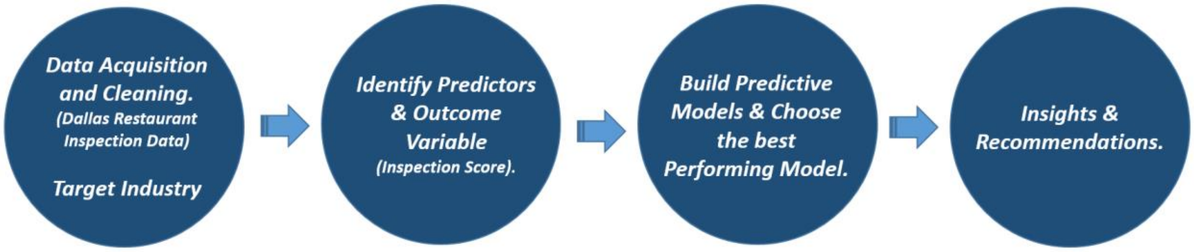
TABLE OF CONTENTS

OVERVIEW OF PROJECT



A Four Stage Project as shown above.
Stage 1: First Stage is mostly about Identifying, acquiring, understanding and cleaning the data. Once completed the target firm is identified.
Stage 2: This involved identifying predictors and outcome variables.
Stage 3: This stage comprises of building different predictive models and choose the best of top 3 high performance models.
Stage 4: This involves Drawing out Insights from the chosen model and come up with actionable business recommendations for the target firm.

BUSINESS UNDERSTANDING

Firm Description

Online review websites are one of the best lead generation tools for restaurants. In today's digital age, diners swiftly become online critics on restaurant review sites. Through their ratings and reviews, they're making personal recommendations not only to their friends and family but also to the whole world. **OpenTable** is one such online restaurant reservation and restaurant review platform. OpenTable was founded in 1998 and have been powering dining experiences for more than 20 years now. The company was founded by Chuck Templeton in San Francisco, after getting frustrated by the need to call multiple restaurants to find the right place to go to dinner.

Restaurants sign up with the services offered by OpenTable and use the company's back-end software to process the reservations through their website or mobile app, resulting in a real-time reservation system for both diners and restaurants.

OpenTable allows users to search for restaurants and reservations based on parameters like dates, times, cuisine, and price range. The customers can also leave reviews and ratings about restaurants in the form of feedback by using a five-point rating system to judge the experience across different categories including Food, Ambience, Value, and Service, which all feed into a restaurant's overall score. Customers can even use the platform to read up on reviews.

As of Q4 2019, since its inception, OpenTable has seated more than 2 billion diners via online reservations, representing more than $91 billion spent at partner restaurants. OpenTable has integrations with more than 600 brands, including Amazon Alexa, Facebook Messenger, Google, Instagram, Snapchat, TripAdvisor, Yahoo! and Zagat. In aggregate, these integrations account for 17 percent of diners seated each month via OpenTable.

OpenTable's workforce consists of around 1200 employees in 10 offices around the globe.

An Interview Given by 'OpenTable' Christa Quarles about the challenges to the business in 2018. Link to the Article

Firm's Geographic footprint

The company is head-quartered in San Francisco, California. Its home market consists of the United States. The company has also expanded its market globally and is active in more than 80 countries, including Australia, Canada, Germany, Ireland, Italy, Japan, Mexico, the Netherlands, Spain, United Kingdom and the United States.

Line of Business for Analysis

The line of business that is the subject of our analysis would be reviews and recommendations. All food establishments are inspected by a licensed professional by Health Services once every 6 months. Scores only represent a snapshot of the facility at the time of inspection. Survey data shows that 94% of U.S. diners are influenced by online reviews, especially when they're on the hunt for somewhere new to eat out. Another survey by Bright Local says that about 60% reported reading a review for a restaurant before they went. The statistics have gone on to reveal a simple fact: restaurants with the most positive reviews on restaurant review websites get the most bookings. The more positive experiences listed, the better. Hence, reviews combined with inspection score of the restaurant have an important role in determining the success or downfall of a restaurant. Our focus will be specific to restaurants in Dallas.

SWOT Analysis

STRENGTHS	WEAKNESSES	OPPORTUNITIES	THREATS
❑ *20+ Years of Industry Experience.* ❑ *Loyal Customers – 60,000 Restaurants and 124 Million Diners Globally.* ❑ *Conversion Based Charge.*	❑ *High Priced – The services offered by Open Table are highly priced with expensive gadgets.* ❑ *Limited Features.*	❑ *Company is Looking for enhanced and unique features to fuel the platform.* ❑ *Existing customer base and brand awareness ease the marketing.*	❑ *Market competitors like Yelp are expanding their customer network rapidly, also constantly scaling up their data.*

Strengths:

Since 1998, OpenTable has dominated the virtual reservation landscape and is considered to be a leading provider of real-time online reservations for diners and guest management solutions for restaurants. The company has a passion for hospitality and helps restaurants grow their business with many of their employees having worked in the service industry.

One benefit to advertising on OpenTable, unlike other search engines, OpenTable does not charge on a click or impression basis, but only if you seat a guest and that person is physically seated in your restaurant.

OpenTable's platform manages 124 million diners on average each month at more than 60,000 restaurants globally. OpenTable also finds restaurants for more than 31 million diners each month via online reservations. OpenTable diners write more than 1 million restaurant reviews every month.

Weaknesses:

Over the years OpenTable has tested and rolled out a number of different tactics to retain and build their audience, but at the end of the day pricing continues to be the main decider. OpenTable is the most expensive option for restaurants that take reservations, with $249 monthly fees + a seated cover charge of 25 ¢ to $1 depending on how the table is booked. OpenTable's main use case is to make reservations, and the reviews are a secondary feature.

Opportunities:

To compensate with their high prices, OpenTable can offer more services to restaurants to improve their businesses. The company is looking to leverage its scale and experience in the space to become a data-fueled recommendation engine for diners and a source of quality intelligence for restaurant operators.

The data and actionable insights can help OpenTable provide more control and flexibility to optimize and increase revenue, while also helping to ensure a positive experience for customers.

Threats:

OpenTable has lost ground to rivals over the last few years. OpenTable can clearly feel competitors nipping at its heels: As CEO Christa Quarles told Skift Table in her own recent interview, "We've got to keep pace with the challenges that our restaurants feel so that we can be a preferred partner." Yelp is continuing its evolution from a user-generated review platform to meeting the needs of diners who want to plan ahead, walk in for seating, or get food delivered. According to internal data, Yelp managed 22 million diners in December 2019, with bookings directly through the Yelp app showing a threefold increase year-over-year for the fourth quarter. The December figure of 22 million diners managed is a shot across the bow of OpenTable, which has dominated the virtual reservation landscape since 1998. Yelp provides routine inspection information of restaurants obtained from HDScores, while HDScores collects public data directly from local health department.

DATA ANALYSIS AND UNDERSTANDING

Data Description

The dataset chosen for the term project was acquired from Dallas Open Data Platform. Dallas Open Data is an invaluable resource for anyone to easily access data published by the City.
Source

Link : https://www.dallasopendata.com/City-Services/Restaurant-and-Food-Establishment-Inspections-Octo/dri5-wcct

About the Data

- This data set is intended to communicate the name of establishment, the physical location of the establishment, the date the inspection was conducted, the overall score for the inspection, and the point deduction for the individual violations.
- The actual data set has ***44,000*** rows and ***114*** columns. After the process of data cleaning came up with ***43,880*** rows and 21 columns.
- Time frame of the data - The dataset captures data from October 2016 to Present.

Each row in the dataset represents a ***Facility Inspection.***

Cleaning the source data

Violation Description Columns in Source file contained 50 different violations which are grouped into 11 categories as follows.

- ✓ **Facility_Cleanliness**
- ✓ **Food_Temperature**
- ✓ **Food_Contamination**

- ✓ **Worker_Cleanliness**
- ✓ **Hazardous_Chemicals**
- ✓ **Facility_Layout**
- ✓ **Certification**
- ✓ **Documentation**
- ✓ **Equipment**
- ✓ **Facility_Amenities**
- ✓ **Other**

Violation Description - 1
*27 Cooling, heating, and holding capacities. Equipment
*27 Cool TCS foods using proper methods
*27 Cooling, heating, and holding capacities. Equipment
*27 Cooling, heating, and holding capacities. Equipment
*27 Cooling method, criteria - smaller portions
*27 Cooling, heating, and holding capacities. Equipment
*27 Cooling, heating, and holding capacities. Equipment
*27 Cooling method, criteria - placing the food in shallow pans
*27 Cooling, heating, and holding capacities. Equipment
*27 Cooling, heating, and holding capacities. Equipment
*27 Cooling, heating, and holding capacities. Equipment
*27 Cooling, heating, and holding capacities. Equipment

Above descriptions can be put into the category Food Temperature using keyword ‘cooling’. Likewise other descriptions are categorized based on the keywords.

Data Dictionary

Column Name	Derived	Description	Parent Column
Restaurant.Name	No	Name of the Facility being Inspected	NA
Inspection.Type	No	Code indicating the inspection type, such as Routine, Follow-up, Complaint, Temporary and Mobile. • Routine Inspections – are conducted at least once every six months • Follow-up Inspections – are conducted as a result of poor sanitation issues, low scores • Complaints Inspections – General Sanitation/Hygienic Practices /Illness Investigation, Smoking and Other	NA
Quality_rating	Yes	Holds values 1 through 5. • 100-90 (Meets Consumer Health Division standards (Excellent Quality) -5 • 89-80 (Meets Consumer Health Division standards (Good Quality) -4 • 79-70 (Requires follow-up inspection within 30 days)-3 • 69-60 (Requires follow-up inspection within 10 days)-2 • 59 and below (Requires follow-up inspection within 24 hours)- 1	Inspection.Score
Inspection.Score	No	The aggregate score from the inspection violations. Please note not every violation will reflect a point deduction as establishments are allowed to correct violations during the inspection process, and therefore no reduction in the overall score is reflected for the violation.	NA
Recommendation	Yes	Holds the value recommended or Not Recommended based on Quality Rating. • Quality_ Rating 5 & 4 is Recommended. • Quality_rating 1,2,3 is Not Recommended.	Quality_rating

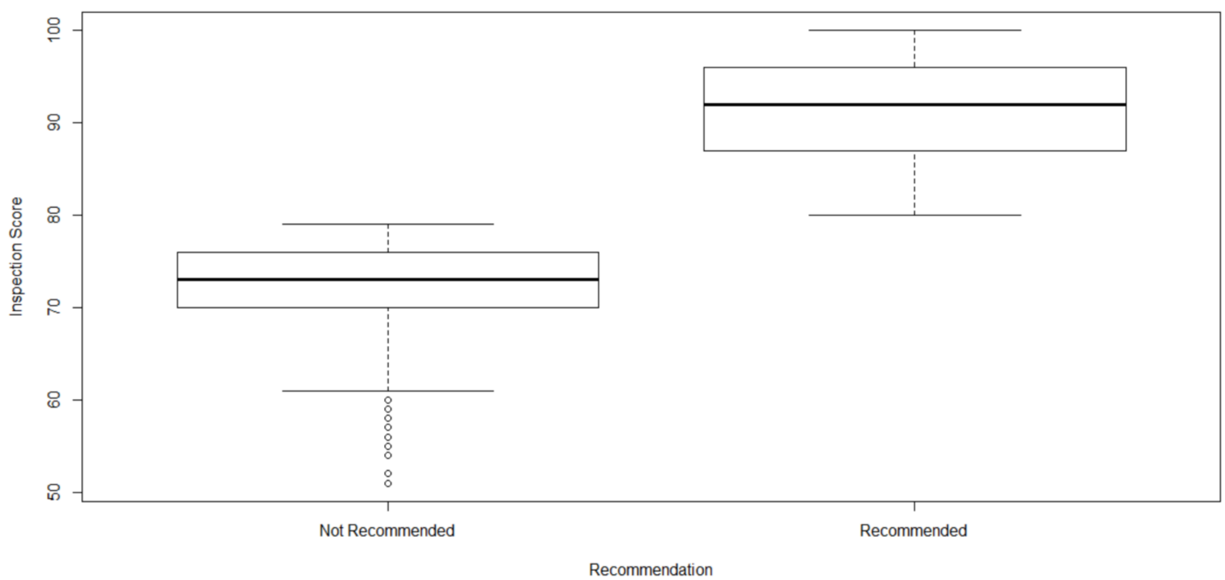
Facility_Cleanliness	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occurred e.g.. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurrence of category violation	Violation Description #
Food_Temperature	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occurred e.g.. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurrence of category violation	Violation Description #
Food_Contamination	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occurred e.g.. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurrence of category violation	Violation Description #
Worker_Cleanliness	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occurred e.g.. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurrence of category violation	Violation Description #
Hazardous Chemicals	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occured eg. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurance of category violation	Violation Description #
Facility_Layout	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occured eg. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurance of category violation	Violation Description #
Certification	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occured eg. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurance of category violation	Violation Description #
Documentation	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occured eg. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurance of category violation	Violation Description #
Equipment	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occured eg. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurance of category violation	Violation Description #
Facility_Amenities	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occured eg. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurance of category violation	Violation Description #
Other	Yes	Represents the violation category that indicates the number of rules the facility has violated under that category. Holds a whole number value based on number of times Violation occured eg. 0 for 0 Violations specific to the category, 2 for 2 Violations specific to the category <u>Derivation</u> If Parent Column value is corresponds to a category a counter is incremented with each occurance of category violation	Violation Description #
Inspection.Month	No	The month in which the Inspection has taken place	NA
Inspection.Year	No	The year In which the inspection has taken place	NA
Lat.Long.Location	No	The latitude and longitude of the facility	NA
Zip.Code	No	Area zip code of the facility	NA
Season	Yes	Indicating the season (Summer, Fall, Winter, Spring) in which the inspection took place.	Inspection Month

Summarize the Dataset

Summary of Numerical Variables

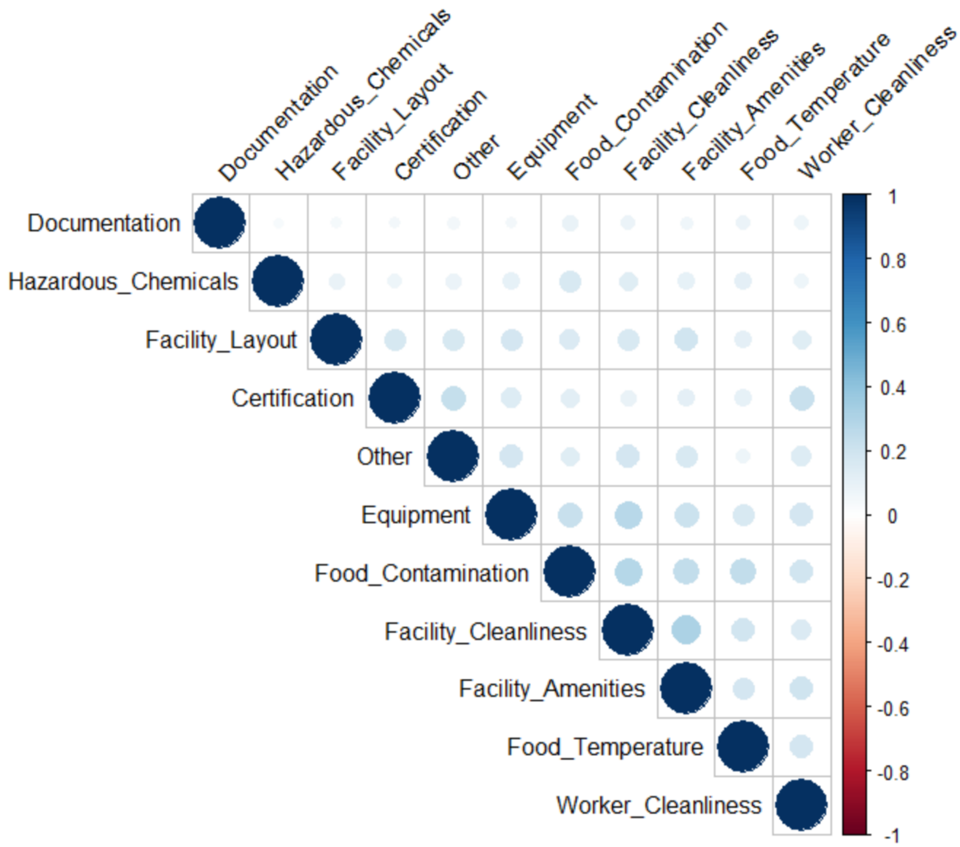
	Minimum	Maximum	Median	Mean	Standard_Deviation	Number_of_Missing_Values
Inspection.Score	51	100	91	90.35072926	7.0451596	0
Facility_Cleanliness	0	5	1	0.86267092	0.9265323	0
Food_Temperature	0	6	0	0.31342297	0.6252599	0
Food_Contamination	0	6	1	0.86561076	1.0330630	0
Worker_Cleanliness	0	5	0	0.34888332	0.6190593	0
Hazardous_Chemicals	0	3	0	0.16595260	0.3751534	0
Facility_Layout	0	6	0	0.50444394	0.7120150	0
Certification	0	4	0	0.34908842	0.5167418	0
Documentation	0	3	0	0.03999544	0.2039296	0
Equipment	0	5	0	0.37470374	0.6261819	0
Facility_Amenities	0	8	1	0.89735643	1.0109780	0
Other	0	4	0	0.39086144	0.5744814	0

Side by side box plot of Inspection Score to recommendation.



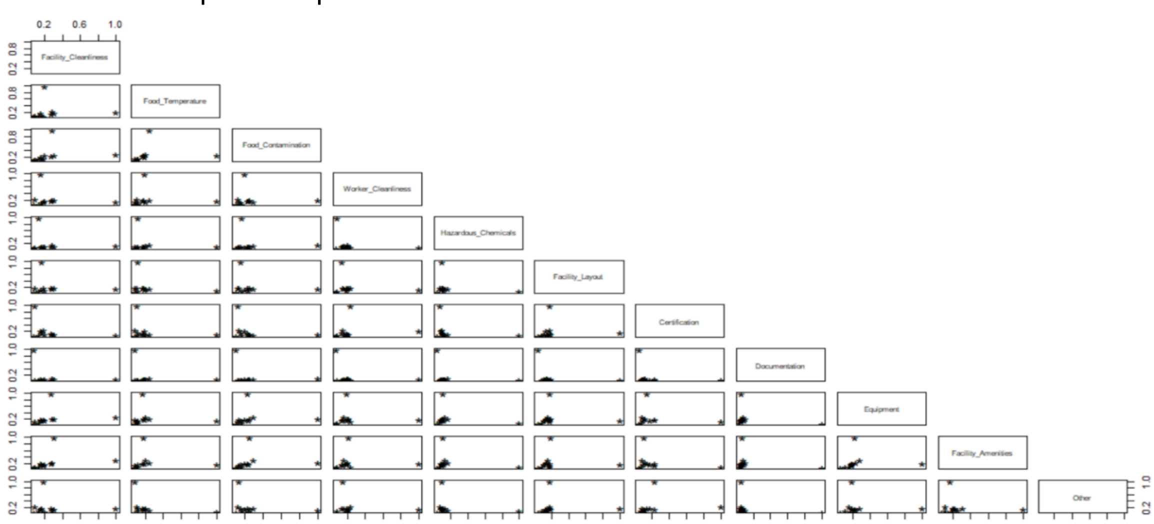
The above box plot shows the distribution of inspection score to the correlated derived variable recommendation in dataset.

Correlation matrix of predictors

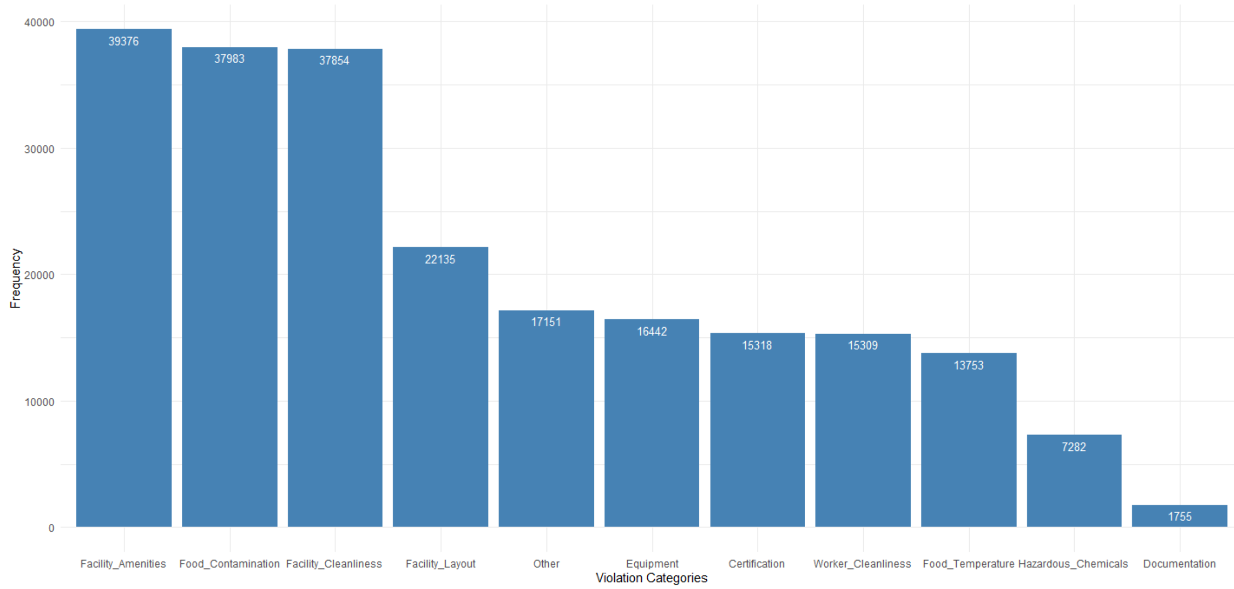


All the variables lie within the value 0 - 0.3 which gives a lower correlation which shows these are good predictors.

Pairwise scatter plots for predictors



Bar chart for the frequency of the occurrences of violations in different violation categories.



DATA MODELING

Goals and Objectives

By introducing our models, OpenTable enables Restaurants to be better prepared for inspection through the prediction of outcome variable 'Inspection Score' and help the restaurant owners to know the desirability of their restaurant through the next target variable 'Recommendation'.
The target variable 'Inspection Score' for our linear model holds an integer value between 51 and 100 and tells whether the restaurant adheres to the rules put out by Dallas Inspection Authority.
The target variable for our logistic regression model, 'Recommendation', holds categorical values: 'Recommend' and 'Not Recommend'. This variable lets the restaurant owner know whether the restaurant will be recommended or not.

Build the Models

Model 1: Full Linear Regression Model

Target Variable: Inspection.Score

Predictors: All the remaining variables (Inspection.Type, Zip.Code, Season, Facility_Cleanliness, Food_Temperature, Food_Contamination, Worker_Cleanliness, Hazardous_Chemicals, Facility_Layout, Certification, Documentation, Equipment, Facility_Amenities, Other, Inspection.Year)

```
> # Creating a full linear regression model.
> iv.lm <- lm(Inspection.Score ~ . , data = train.df )
> options(scipen=999)
> summary(iv.lm)

Call:
lm(formula = Inspection.Score ~ ., data = train.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.3792  -0.8213   0.2274   0.8346  17.9774 

Coefficients:
                          Estimate  Std. Error  t value              Pr(>|t|)    
(Intercept)            -65.5302674  28.8038253   -2.275                0.0229 *  
Inspection.TypeFollow-up -0.2278414  0.3575699   -0.637                0.5240    
Inspection.TypeRoutine  -0.0554895   0.3511441   -0.158                0.8744    
Zip.Code                 0.0021932   0.0003829    5.728          0.0000000102 ***
SeasonSpring             0.0431393   0.0236873    1.821                0.0686 .  
SeasonSummer             0.0021614   0.0246633    0.088                0.9302    
SeasonWinter            -0.0060442   0.0222079   -0.272                0.7855    
Facility_Cleanliness    -1.6640422   0.0099588 -167.093 < 0.0000000000000002 ***
Food_Temperature        -2.7770293   0.0140800 -197.232 < 0.0000000000000002 ***
Food_Contamination      -2.0277400   0.0088458 -229.232 < 0.0000000000000002 ***
Worker_Cleanliness      -2.6191408   0.0143489 -182.533 < 0.0000000000000002 ***
Hazardous_Chemicals     -2.9620182   0.0224925 -131.689 < 0.0000000000000002 ***
Facility_Layout         -2.2563497   0.0121834 -185.198 < 0.0000000000000002 ***
Certification           -1.9899118   0.0169795 -117.195 < 0.0000000000000002 ***
Documentation           -2.2498863   0.0412071  -54.599 < 0.0000000000000002 ***
Equipment               -1.7098821   0.0143302 -119.320 < 0.0000000000000002 ***
Facility_Amenities      -0.9044813   0.0089796 -100.726 < 0.0000000000000002 ***
Other                   -0.9190339   0.0152548  -60.245 < 0.0000000000000002 ***
Inspection.YearFY2018    0.1793603   0.0209608    8.557 < 0.0000000000000002 ***
Inspection.YearFY2019    0.2556672   0.0222175   11.507 < 0.0000000000000002 ***
Inspection.YearFY2020    0.2957603   0.0301124    9.822 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.446 on 30694 degrees of freedom
Multiple R-squared:  0.958,	Adjusted R-squared:  0.9579 
F-statistic: 3.497e+04 on 20 and 30694 DF,  p-value: < 0.00000000000000022
```

Equation of the fitted model:

Algebraically, the equation for multiple linear regression is:

$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots\ldots + \beta pxp + \varepsilon$

where $\varepsilon \sim N(0, \sigma 2)$

β_0 = Estimate value in the (Intercept) row = -65.5302674

β_1 = Estimate value in the Inspection.TypeFollow-up row= -0.2278414

β_2 = Estimate value in the Inspection.TypeRoutine row = -0.0554895

β_3 = Estimate value in the Zip.Code row = 0.0021932

β_4 = Estimate value in the SeasonSpring row = 0.0431393

β_5 = Estimate value in the SeasonSummer row = 0.0021614

β_6 = Estimate value in the SeasonWinter row = -0.0060442

β_7 = Estimate value in the Facility_Cleanliness row = -1.6640422

β_8 = Estimate value in the Food_Temperature row = -2.7770293

β_9 = Estimate value in the Food_Contamination row = -2.0277400

β_{10} = Estimate value in the Worker_Cleanliness row = -2.6191408

β_{11} = Estimate value in the Hazardous_Chemicals row = -2.9620182

β_{12} = Estimate value in the Facility_Layout row = -2.2563497

β_{13} = Estimate value in the Certification row = -1.9899118

β_{14} = Estimate value in the Documentation row = -2.2498863

β_{15} = Estimate value in the Equipment row = -1.7098821

β_{16} = Estimate value in the Facility_Amenities row = -0.9044813

β_{17} = Estimate value in the Other row = -0.9190339
β_{18} = Estimate value in the Inspection.YearFY2018 row = 0.1793603
β_{19} = Estimate value in the Inspection.YearFY2019 row = 0.2556672
β_{20} = Estimate value in the Inspection.YearFY2020 row = 0.2957603

σ = the Residual standard error = 1.446

Plugging these in above equation yields:
$Y = (-65.5302674) + [(-0.2278414)(x_1) \text{ OR } (-0.0554895)(x_2)] + (0.0021932)(x_3) + [(0.0431393)(x_4) \text{ OR } (0.0021614)(x_5) \text{ OR } (-0.0060442)(x_6)] + (-1.6640422)(x_7) + (-2.7770293)(x_8) + (-2.0277400)(x_9) + (-2.6191408)(x_{10}) + (-2.9620182)(x_{11}) + (-2.2563497)(x_{12}) + (-1.9899118)(x_{13}) + (-2.2498863)(x_{14}) + (-1.7098821)(x_{15}) + (-0.9044813)(x_{16}) + (-0.9190339)(x_{17}) + [(0.1793603)(x_{18}) \text{ OR } (0.2556672)(x_{19}) \text{ OR } (0.2957603)(x_{20})] + \varepsilon$, where $\varepsilon \sim N(0, (1.446)^2)$

Y = (-65.5302674) + [(-0.2278414)(Inspection.TypeFollow-up) OR (-0.0554895)(Inspection.TypeRoutine)] + (0.0021932)(Zip.Code) + [(0.0431393)(SeasonSpring) OR (0.0021614)(SeasonSummer) OR (-0.0060442)(SeasonWinter)] + (-1.6640422)(Facility_Cleanliness) + (-2.7770293)(Food_Temperature) + (-2.0277400)(Food_Contamination) + (-2.6191408)(Worker_Cleanliness) + (-2.9620182)(Hazardous_Chemicals) + (-2.2563497)(Facility_Layout) + (-1.9899118)(Certification) + (-2.2498863)(Documentation) + (-1.7098821)(Equipment) + (-0.9044813)(Facility_Amenities) + (-0.9190339)(Other) + [(0.1793603)(Inspection.YearFY2018) OR (0.2556672)(Inspection.YearFY2019) OR (0.2957603)(Inspection.YearFY2020)] + ε, where $\varepsilon \sim N(0, (1.446)^2)$

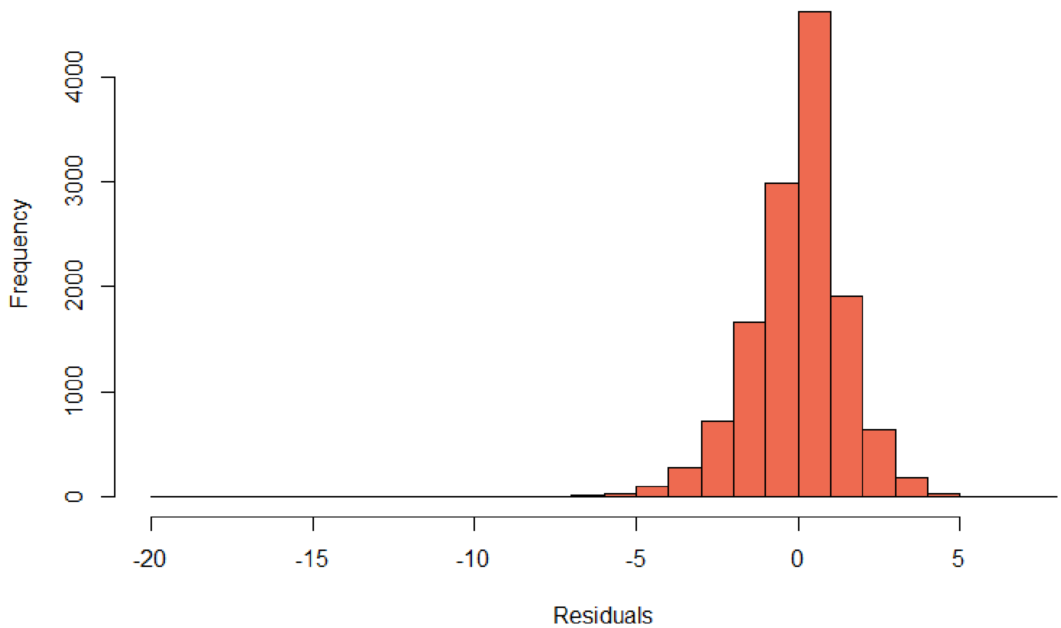
Adjusted R-squared: 0.9579

Model accuracy with Validation data:

```
> predicteddata <- predict(iv.lm, valid.df )
> accuracy(round(predicteddata),valid.df$Inspection.Score)
                  ME     RMSE      MAE         MPE     MAPE
Test set -0.03334599 1.510239 1.057729 -0.06169788 1.209669
```

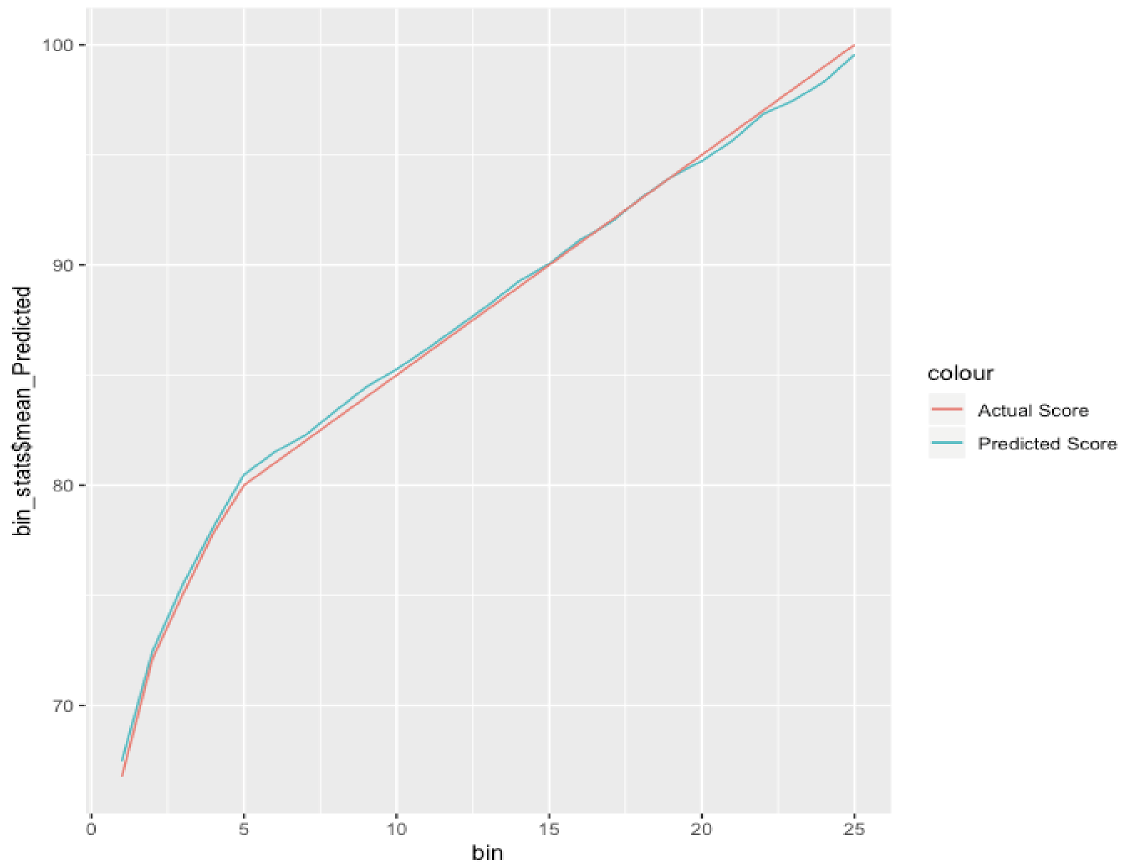
Mean Error: -0.3334599
Root Mean Squared Error: 1.510239
Mean Absolute Error: 1.057729
Mean Percentage Error: -0.06169788
Mean Absolute Percentage Error: 1.209669

Residual plot of validation data with Full Linear Regression model:



The histogram of the residuals show that most of the errors are between -3 and +3.

Comparison of Actual Vs. Predicted Plot with Validation data:

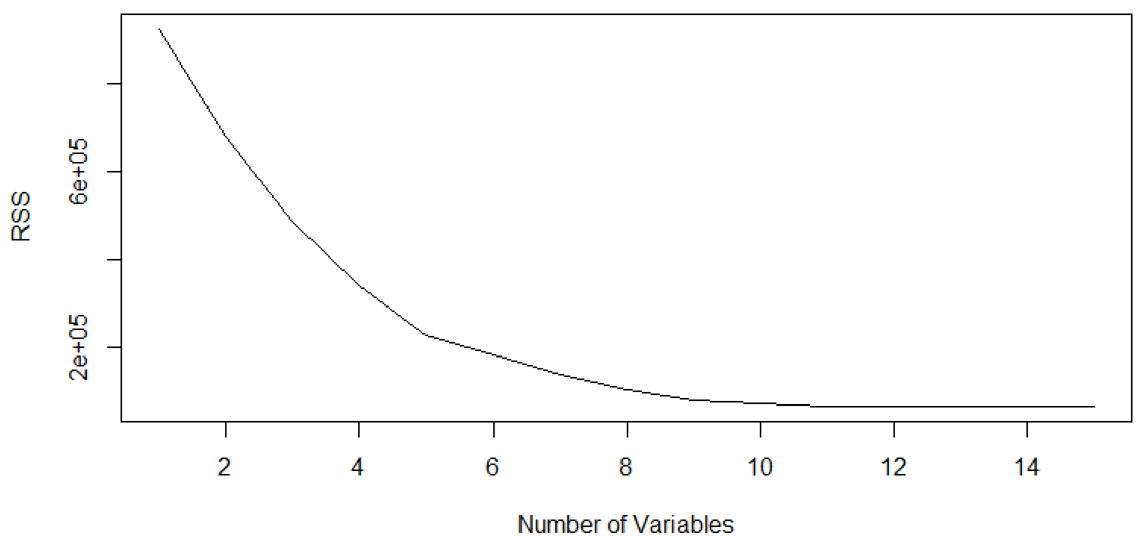


From the above plot, we can infer that there's a strong similarity between the model's predicted score and its actual scores.

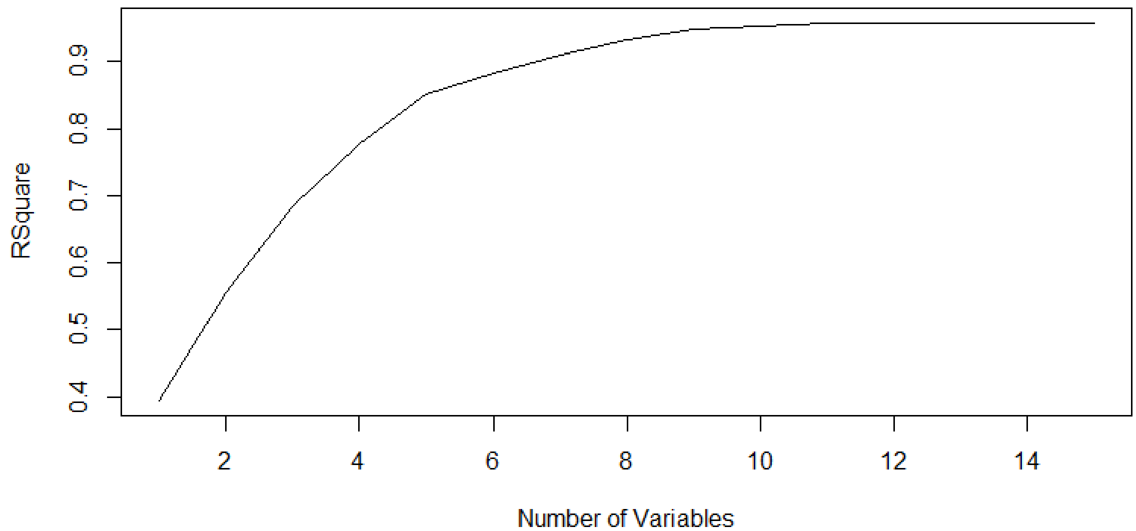
Model 2: Parsimonious Linear Regression Model

Target Variable: Inspection.Score
Best subset selection:

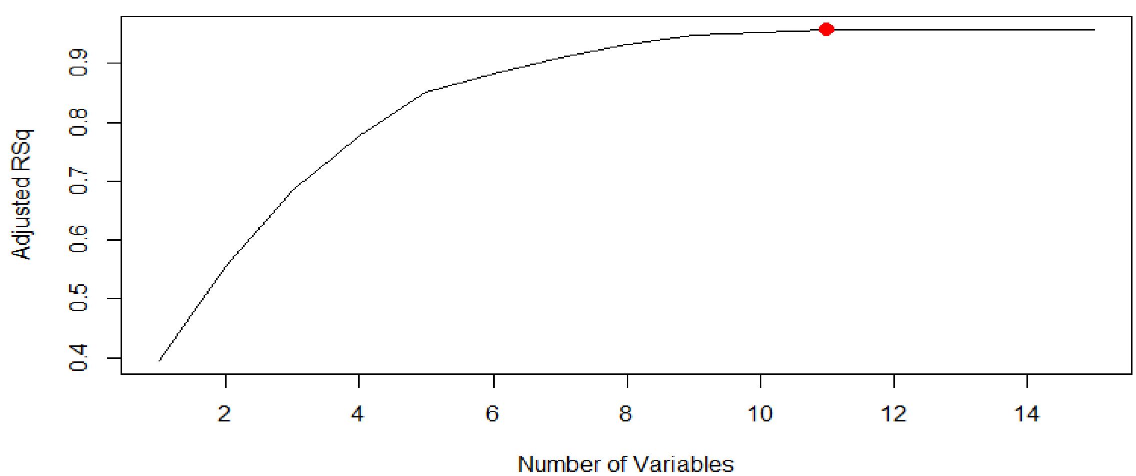
The regsubsets() function (part of the leaps library) performs best subset selection by identifying the best model that contains a given number of predictors. So, I use regsubsets() in package leaps to run an **exhaustive** search. The regsubsets() will calculate adjusted r2 for every possible combination of predictors which we plot on a graph.



RSS: For Residual Sum of Squares, which we want to minimize, so if we include more than 10 variables, RSS will be low.

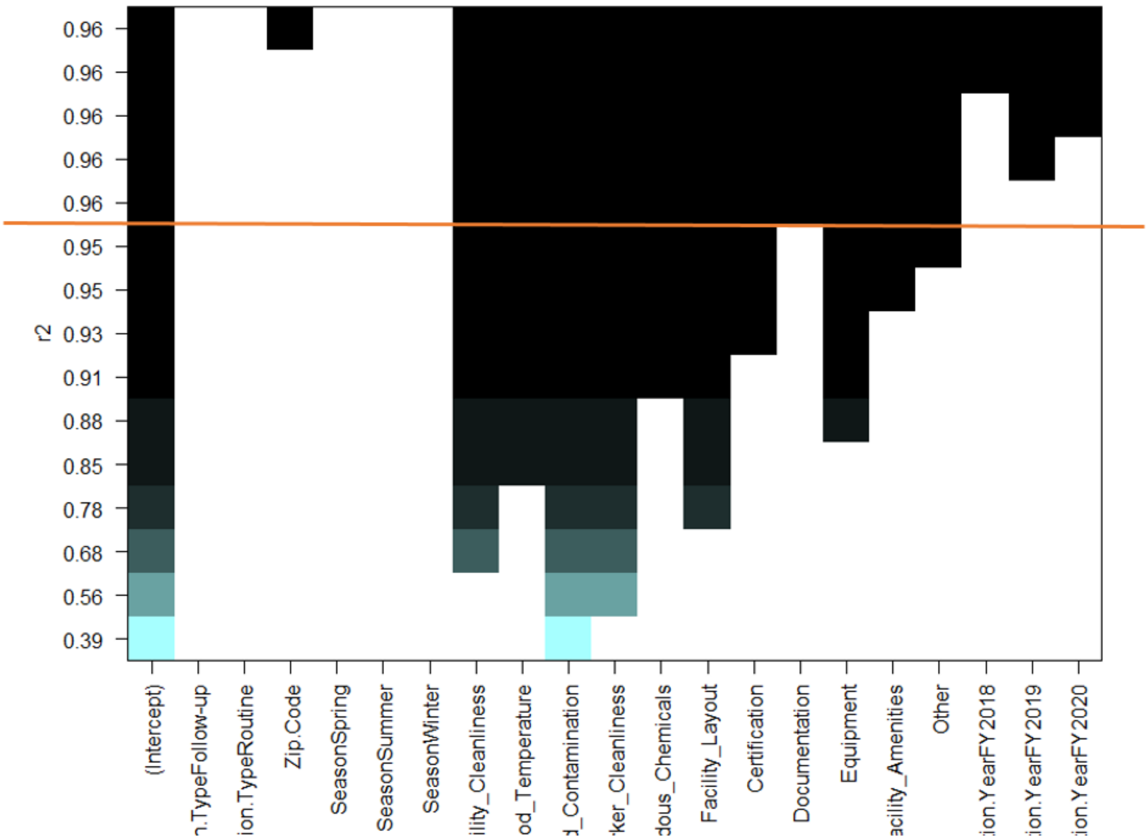


RSQ: When we include about more than 10 variables, we reach the highest r-square we can get. About 95%.



Adjusted R-sq: We see that 11 variables have the highest Adjusted Rsquare

Based on the above graphs I have chosen the top 11 significant predictors.



Based on the above plot, we choose our predictors as:

Predictors: Food_Contamination, Worker_Cleanliness, Facility_Cleanliness, Facility_Layout, Food_Temperature, Equipment, Hazardous_Chemicals, Certification, Facility_Amenities, Other, Documentation

```
> # Building a parsimonious model with 11 best significant variables which are the violation variables.
> pariv.lm <- lm(Inspection.Score ~ Food_Contamination + Worker_Cleanliness + Facility_Cleanliness +
+                  Facility_Layout + Food_Temperature + Equipment + Hazardous_Chemicals + Certification +
+                  Facility_Amenities + Other + Documentation , data = train.df )
> options(scipen=999)
> summary(pariv.lm)

Call:
lm(formula = Inspection.Score ~ Food_Contamination + Worker_Cleanliness +
    Facility_Cleanliness + Facility_Layout + Food_Temperature +
    Equipment + Hazardous_Chemicals + Certification + Facility_Amenities +
    Other + Documentation, data = train.df)

Residuals:
     Min       1Q   Median       3Q      Max
-19.5519  -0.8535   0.2549   0.8330  18.0552

Coefficients:
                      Estimate Std. Error t value            Pr(>|t|)
(Intercept)          99.551880   0.014382 6922.07 <0.0000000000000002 ***
Food_Contamination   -2.026225   0.008846 -229.05 <0.0000000000000002 ***
Worker_Cleanliness   -2.631030   0.014360 -183.22 <0.0000000000000002 ***
Facility_Cleanliness -1.656313   0.009958 -166.33 <0.0000000000000002 ***
Facility_Layout      -2.257070   0.012213 -184.81 <0.0000000000000002 ***
Food_Temperature     -2.777986   0.014099 -197.04 <0.0000000000000002 ***
Equipment            -1.720505   0.014343 -119.95 <0.0000000000000002 ***
Hazardous_Chemicals  -2.975096   0.022546 -131.96 <0.0000000000000002 ***
Certification        -1.975842   0.016994 -116.27 <0.0000000000000002 ***
Facility_Amenities   -0.906309   0.009002 -100.67 <0.0000000000000002 ***
Other                -0.911542   0.015275  -59.67 <0.0000000000000002 ***
Documentation        -2.260421   0.041232  -54.82 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.451 on 30703 degrees of freedom
Multiple R-squared:  0.9577,    Adjusted R-squared:  0.9577
F-statistic: 6.315e+04 on 11 and 30703 DF,  p-value: < 0.00000000000000022
```

Equation of the fitted model:

Algebraically, the equation for multiple linear regression is:

Y = β0 + β1x1 + β2x2 + …… + βpxp + ε

where ε ~ N(0, σ2)

β_0 = Estimate value in the (Intercept) row = 99.55188

β_1 = Estimate value in the Food_Contamination row = -2.026225

β_2 = Estimate value in the Worker_Cleanliness row = -2.63103

β_3 = Estimate value in the Facility_Cleanliness row = -1.656313

β_4 = Estimate value in the Facility_Layout row = -2.25707

β_5 = Estimate value in the Food_Temperature row = -2.777986

β_6 = Estimate value in the Equipment row = -1.720505

β_7 = Estimate value in the Hazardous_Chemicals row = -2.975096

β_8 = Estimate value in the Certification row = -1.975842

β_9 = Estimate value in the Facility_Amenities row = -0.906309

β_{10} = Estimate value in the Other row = -0.911542

β_{11} = Estimate value in the Documentation row = -2.260421

σ = the Residual standard error = 1.451

Plugging these in above equation yields:

$Y = (99.55188) + (-2.026225)(x_1) + (-2.63103)(x_2) + (-1.656313)(x_3) + (-2.25707)(x_4) + (-2.777986)(x_5) + (-1.720505)(x_6) + (-2.975096)(x_7) + (-1.975842)(x_8) + (-0.906309)(x_9) + (-0.911542)(x_{10}) + (-2.260421)(x_{11}) + \varepsilon$, where $\varepsilon \sim N(0, (1.451)^2)$

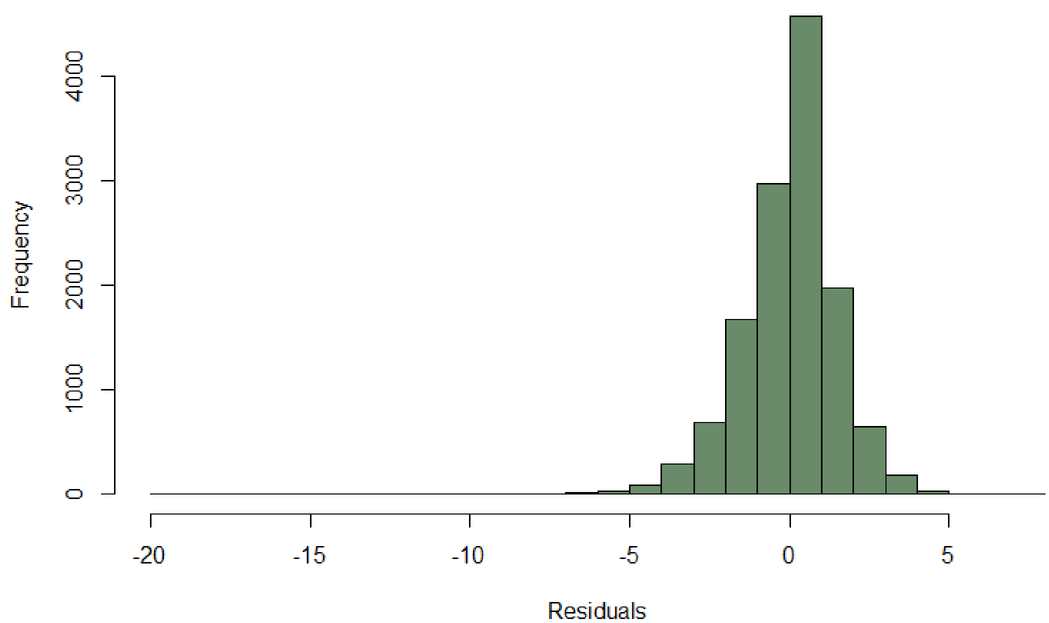
Y = (99.55188) + (-2.026225)(Food_Contamination) + (-2.63103)(Worker_Cleanliness) + (-1.656313)(Facility_Cleanliness) + (-2.25707)(Facility_Layout) + (-2.777986)(Food_Temperature) + (-1.720505)(Equipment) + (-2.975096)(Hazardous_Chemicals) + (-1.975842)(Certification) + (-0.906309)(Facility_Amenities) + (-0.911542)(Other) + (-2.260421)(Documentation) + ε, where ε ~ N(0, $(1.451)^2$)

Adjusted R-squared: 0.9577
Model accuracy with Validation data:

```
> parpredicteddata <- predict(pariv.lm, valid.df )
> accuracy(round(parpredicteddata),valid.df$Inspection.Score)
                  ME     RMSE      MAE         MPE     MAPE
Test set -0.06836308 1.514182 1.046259 -0.09691329 1.199603
```

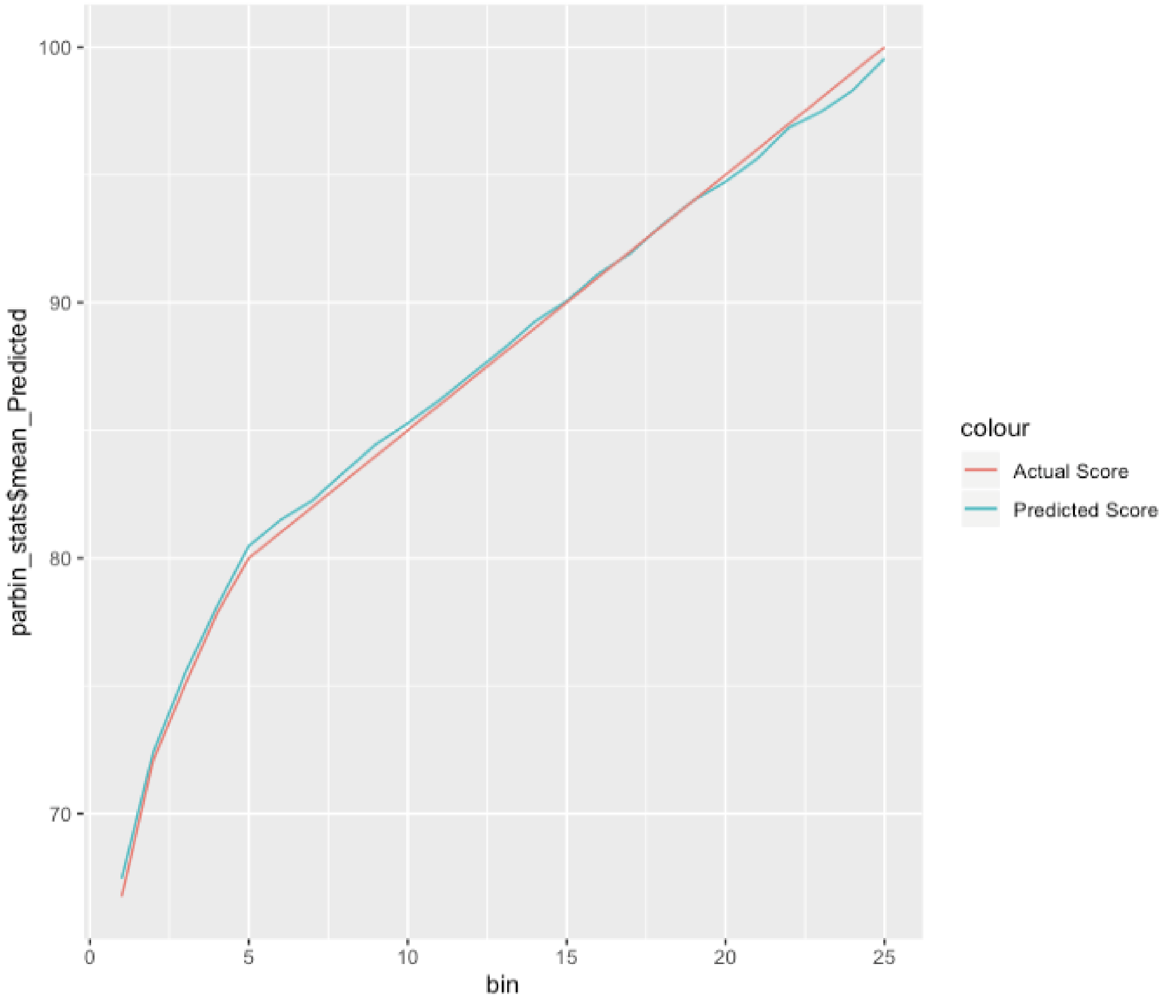
Mean Error: -0.06836308
Root Mean Squared Error: 1.514182
Mean Absolute Error: 1.046259
Mean Percentage Error: -0.09691329
Mean Absolute Percentage Error: 1.199603

Residual plot of validation data with Parsimonious Linear Regression model:



The histogram of the residuals shows that most of the errors are between -3 and +3.

Comparison of Actual Vs. Predicted Plot with Validation data:



Like the comparison of actual vs. predicted plot for the full linear regression model, we can see that there's a strong similarity between the model's predicted score and its actual scores.

Model 3: Logistic Regression Model

Outcome Variable: Recommendation
Predictors: Inspection.Type, Zip.Code, Season, Facility_Cleanliness, Food_Temperature, Food_Contamination, Worker_Cleanliness, Hazardous_Chemicals, Facility_Layout, Certification, Documentation, Equipment, Facility_Amenities, Other, Inspection.Year

```
> iv.glm <- glm(Recommendation ~ . , data = train.df, family = "binomial")
> options(scipen=999)
> summary(iv.glm)

Call:
glm(formula = Recommendation ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0238   0.0001   0.0007   0.0100   2.7068

Coefficients:
                            Estimate  Std. Error z value             Pr(>|z|)
(Intercept)              -397.071476  213.059714  -1.864              0.06237 .
Inspection.TypeFollow-up    2.208629    1.333383   1.656              0.09764 .
Inspection.TypeRoutine      1.831078    1.300248   1.408              0.15906
Zip.Code                    0.005548    0.002833   1.959              0.05015 .
SeasonSpring                0.488510    0.174213   2.804              0.00505 **
SeasonSummer                0.392082    0.178107   2.201              0.02771 *
SeasonWinter                0.032392    0.164868   0.196              0.84424
Facility_Cleanliness       -1.654380    0.077465 -21.356 < 0.0000000000000002 ***
Food_Temperature           -2.952051    0.109824 -26.880 < 0.0000000000000002 ***
Food_Contamination         -2.016983    0.079246 -25.452 < 0.0000000000000002 ***
Worker_Cleanliness         -2.576035    0.106091 -24.281 < 0.0000000000000002 ***
Hazardous_Chemicals        -2.702102    0.147328 -18.341 < 0.0000000000000002 ***
Facility_Layout            -2.395420    0.097986 -24.447 < 0.0000000000000002 ***
Certification              -2.233642    0.125356 -17.818 < 0.0000000000000002 ***
Documentation              -2.524156    0.204513 -12.342 < 0.0000000000000002 ***
Equipment                  -1.788239    0.089951 -19.880 < 0.0000000000000002 ***
Facility_Amenities         -0.964874    0.058153 -16.592 < 0.0000000000000002 ***
Other                      -0.838958    0.097249  -8.627 < 0.0000000000000002 ***
Inspection.YearFY2018       0.509879    0.159553   3.196              0.00140 **
Inspection.YearFY2019      -0.043516    0.157632  -0.276              0.78250
Inspection.YearFY2020       1.096639    0.244225   4.490           0.00000711 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11491.6  on 30714  degrees of freedom
Residual deviance:  1930.7  on 30694  degrees of freedom
AIC: 1972.7

Number of Fisher Scoring iterations: 10
```

Equation of the fitted model:

β_0 = Estimate value in the (Intercept) row = -397.07

β_1 = Estimate value in the Inspection.TypeFollow-up row= 2.20863

β_2 = Estimate value in the Inspection.TypeRoutine row = 1.83108

β_3 = Estimate value in the Zip.Code row = 0.00555

β_4 = Estimate value in the SeasonSpring row = 0.48851

β_5 = Estimate value in the SeasonSummer row = 0.39208

β_6 = Estimate value in the SeasonWinter row = 0.03239

β_7 = Estimate value in the Facility_Cleanliness row = -1.6544

β_8 = Estimate value in the Food_Temperature row = -2.9521

β_9 = Estimate value in the Food_Contamination row = -2.017

β_{10} = Estimate value in the Worker_Cleanliness row = -2.576

β_{11} = Estimate value in the Hazardous_Chemicals row = -2.7021

β_{12} = Estimate value in the Facility_Layout row = -2.3954

β_{13} = Estimate value in the Certification row = -2.2336

β_{14} = Estimate value in the Documentation row = -2.5242

β_{15} = Estimate value in the Equipment row = -1.7882
β_{16} = Estimate value in the Facility_Amenities row = -0.9649
β_{17} = Estimate value in the Other row = -0.839
β_{18} = Estimate value in the Inspection.YearFY2018 row = 0.50988
β_{19} = Estimate value in the Inspection.YearFY2019 row = -0.0435
β_{20} = Estimate value in the Inspection.YearFY2020 row = 1.09664

The equation of the fitted model is:
$P = 1 / (1 + e^{-(\beta 0 + \beta 1x1 + \beta 2x2 + \ldots\ldots + \beta qxq)})$
Therefore,
$P = 1 / (1 + e^{-((-397.07) + [(2.20863)(\text{Inspection.TypeFollow-up})\text{ OR }(1.83108)(\text{Inspection.TypeRoutine})] + (0.00555)(\text{Zip.Code}) + [(0.48851)(\text{SeasonSpring})\text{ OR }(0.39208)(\text{SeasonSummer})\text{ OR }(0.03239)(\text{SeasonWinter})] + (-1.6544)(\text{Facility\_Cleanliness}) + (-2.9521)(\text{Food\_Temperature}) + (-2.017)(\text{Food\_Contamination}) + (-2.576)(\text{Worker\_Cleanliness}) + (-2.7021)(\text{Hazardous\_Chemicals}) + (-2.3954)(\text{Facility\_Layout}) + (-2.2336)(\text{Certification}) + (-2.5242)(\text{Documentation}) + (-1.7882)(\text{Equipment}) + (-0.9649)(\text{Facility\_Amenities}) + (-0.839)(\text{Other}) + [(0.50988)(\text{Inspection.YearFY2018})\text{ OR }(-0.0435)(\text{Inspection.YearFY2019})\text{ OR }(1.09664)(\text{Inspection.YearFY2020})])})$

Model accuracy with the validation data with different cutoff values:

Cutoff Value : 0.8

```
                 Reference
Prediction        Not Recommended Recommended
  Not Recommended             509          55
  Recommended                  97       12504

               Accuracy : 0.9885
```

Cutoff Value : 0.6

```
                 Reference
Prediction        Not Recommended Recommended
  Not Recommended             519          72
  Recommended                  87       12487

               Accuracy : 0.9879
```

Cutoff Value : 0.55

```
                 Reference
Prediction        Not Recommended Recommended
  Not Recommended             548         168
  Recommended                  58       12391

               Accuracy : 0.9828
```

I have chosen our cut off value as 0.8

Model accuracy with Validation data with cutoff value > = 0.8:

```
> predicted.data <- predict(iv.glm, valid.df, type = "response")>=0.8
> predicted.data <- factor(ifelse(predicted.data == TRUE, "Recommended", "Not Recommended"))
> confusionMatrix(predicted.data, valid.df$Recommendation)
Confusion Matrix and Statistics

                 Reference
Prediction        Not Recommended Recommended
  Not Recommended             548         168
  Recommended                  58       12391

               Accuracy : 0.9828
                 95% CI : (0.9805, 0.985)
    No Information Rate : 0.954
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.8201

 Mcnemar's Test P-Value : 0.000000000000415

            Sensitivity : 0.90429
            Specificity : 0.98662
         Pos Pred Value : 0.76536
         Neg Pred Value : 0.99534
             Prevalence : 0.04603
         Detection Rate : 0.04163
   Detection Prevalence : 0.05439
      Balanced Accuracy : 0.94546

       'Positive' Class : Not Recommended
```

True Positive: We predicted 'Not Recommended' and it is 'Not Recommended' = 548
False Negative: We predicted 'Not Recommended' and it is 'Recommended' = 168
False Positive: We predicted 'Recommended' and it is 'Not Recommended' = 58
True Negative: We predicted 'Recommended' and it is 'Recommended' = 12391

Accuracy = (True Positive + True Negative)/ total = 12939/13165 = 0.9828

The **positive class** or the class of interest is 'Not Recommended'.
Sensitivity i.e. how apt the model is to detecting events in the positive class = 0.90429
Specificity i.e. how exact the assignment to the positive class = 0.98662

Transformations performed on the dataset

The initial raw data had 44,000 records, which included 118 records with NA values and 2 records with outliers in the inspection scores (for example: Inspection Score = 0 or –5) which was affecting the MAPE. After removing these records, I ended up with 43,880 records.

There were more than 50 violations in the initial dataset which were categorized into 11 violation categories based on their similarities. Later I created attributes with these categories having values of how many times a restaurant is committing a violation in that category.

We also added attributes such as "Recommendation", "Quality Rating" and "Season" to come up with better predictions. Ex: For the Logistic Model, I did some preprocessing on the categorical variable "Recommendation". I transformed "Recommendation" variable into binary variables, that is, "Recommend" = 1 and "Not Recommend" = 0 and predicted whether the restaurant will be recommended to the customers.

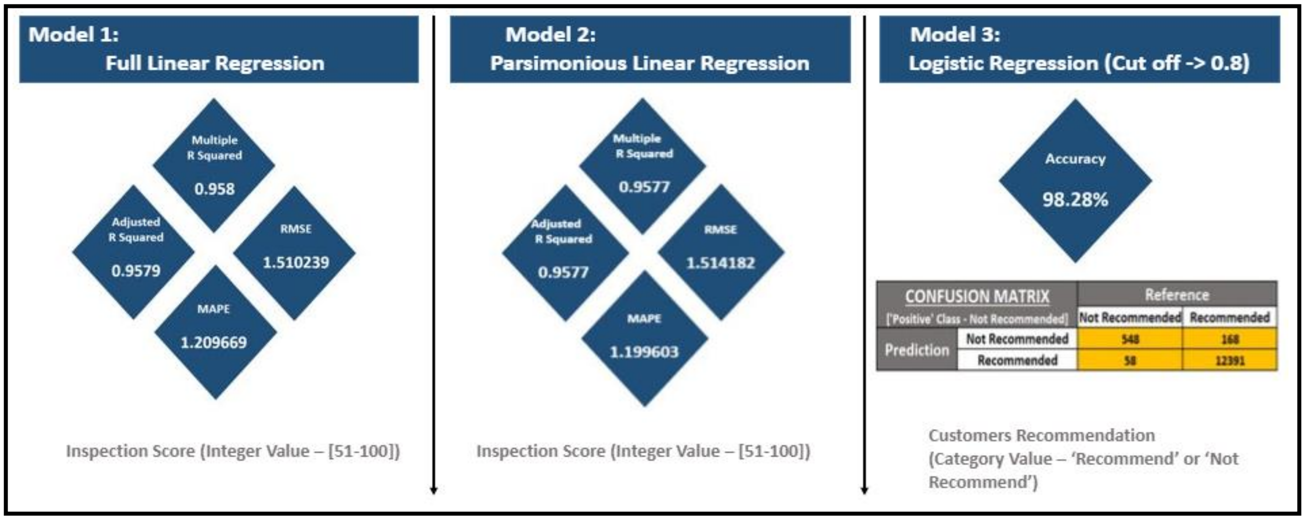
Evaluate Model Performance

The 3 models that I have chosen for our analysis -

- **Full Linear Regression Model** (This model captures the values of all the predictors in the dataset) - The RMSE is 1.510239 which is slightly less and better than the parsimonious model. As the RMSE is low, the accuracy is high. But this model is utilizing all the attributes of the dataset, that can possibly result in ***overfitting*** of the model thereby resulting in better accuracy. This kind of model will fail to better predict the future observations.

- **Parsimonious Linear Regression Model** (This model captures the most significant predictors – all 11 violations) The RMSE is 1.514182 with an adjusted R-square value of 0.9577. The MAPE (1.19) is lower than the MAPE of the previous model.

- **Logistic Regression -** The logistic regression model performs with an accuracy of 98.28% when the cut off is 0.8. The model seems to give even better accuracy when the cut of is decreased to 0.5, but then it would classify all the restaurants as recommended even though in reality they are on the verge of falling into the not recommended category. Therefore, I choose to set a higher cut of percentage.



<u>Model Comparison Sheet</u>

Model
Comparision.xlsx

Chosen Model - Parsimonious Linear Regression Model.

I have chosen Parsimonious Linear Regression Model as our best model. Though the accuracy is slightly lower than Full linear regression model, it predicts with less percentage error when compared to the other model as it only considers the most significant predictors. Opting for this model will also help us to better predict the inspection score for future dataset.

The reason for choosing model 2 over logistic regression model is that the continuous value will give a deeper insight rather than a binary value like 'recommended' or 'not recommended'. An inspection score value also gives an option to restaurant owner to adjust their minimum threshold of inspection score when the feature is implemented.

Limitations of the chosen model

Our dataset comprises fewer records on lower inspection score compared to the higher ones. Having more records with lower inspection score would give us a better predictive model.

RECOMMENDATIONS

Based on our final model following are the recommendations that I'd make to our target firm OpenTable:

- **A paid self-inspection feature for the restaurant:** *Restaurants will have the option to self-check the inspection score and quality rating before inspection happens and always maintain the standard. **This enables Restaurants to be better prepared for inspection.***

- **"Is your restaurant recommended?" Feature:** *A feature which shows whether restaurant is recommended or not recommended based on inspection score. This also allows restaurant owners to adjust the base inspection score above the minimum standard level.*

- ***Top restaurant Chart Feature*** *–Featured list of top quality restaurants based on predicted inspection Score.*

R CODES

Data Cleaning

project
datacleaning.R

Data Understanding

Data_understanding.
r

Multiple Linear Regression – Model 1 & 2

project-multiple-reg
ression.R

Logistic Regression - Model 3

project-logistic-regr
ession-clean.R