AIRBNB & MARKETING

DATA MINING FINAL PRESENTATION

by Kunth Shah

Property Sales Dataset

Independent Variables

- Bedrooms
- Bathrooms
- guests
- openness
- occupancy
- nightly rate
- lead time
- length stay
- Hot tub
- Pool

Dependent Variables

Revenue

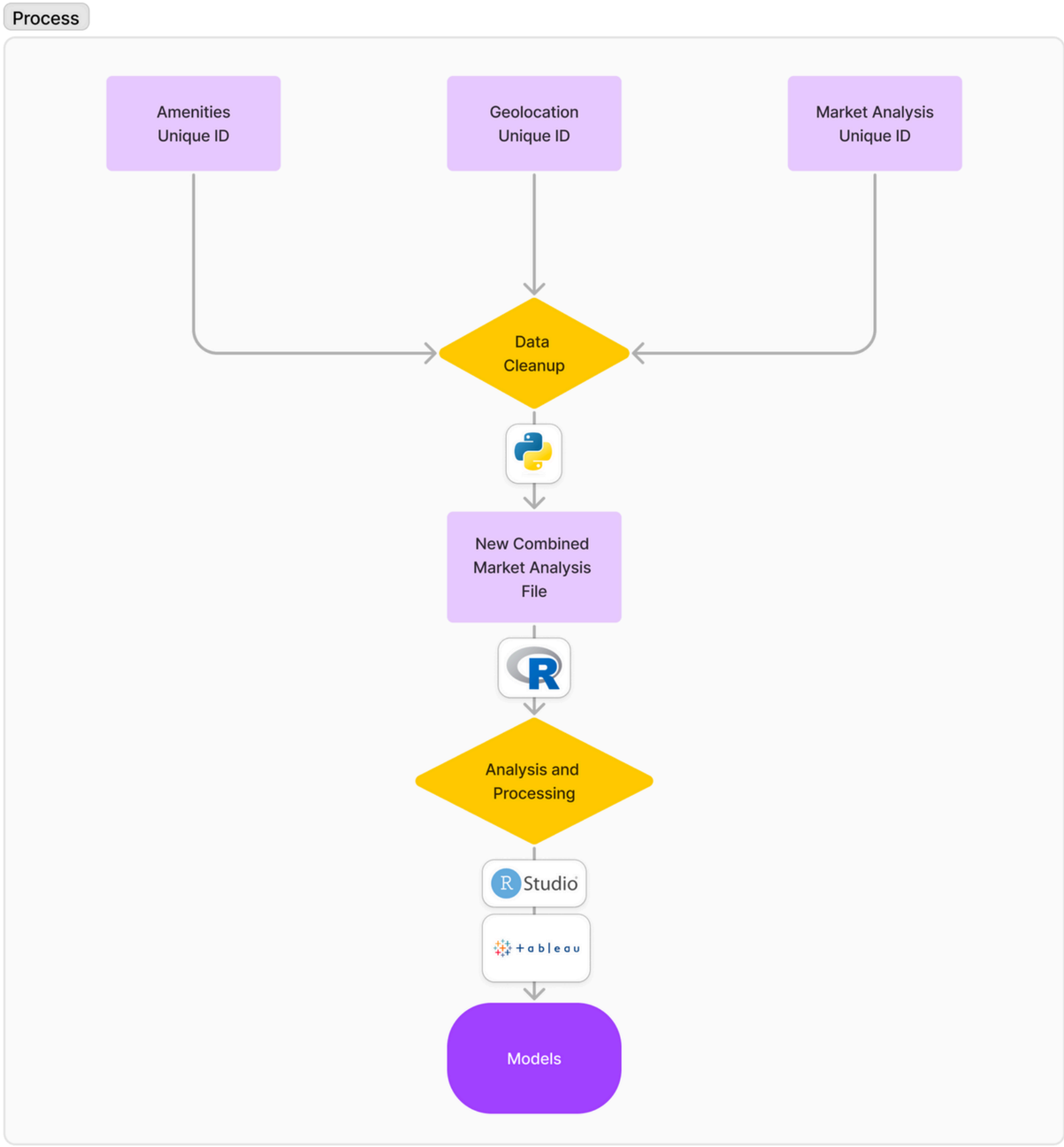
Dataset: https://www.kaggle.com/datasets/computingvictor/zillow-market-analysis-and-real-estate-sales-data

INTRODUCTION

PROCESS AND APPROACH

Tools Used:

Tableau -> Exploratory Research
R -> Analysis and Linear Regression Modelling
Python -> Data Cleaning



INTRODUCTION
DATA CLEANING

1. Amenities.csv
   - Duplicate Data

2. Geo Location.csv
   - Duplicate Data
   - Missing values in street name
   - Values in Latitude and Longitude separated by comma (‘,’), instead of a full-stop (‘.’).

3. Market Analysis 2019.csv
   - Missing Values
   - Values in Revenue are also separated by comma (‘,’), instead of a full-stop (‘.’).

```python
import csv

def replace_commas(input_file, output_file):
    with open(input_file, 'r') as infile, open(output_file, 'w', newline='') as outfile:
        reader = csv.reader(infile, delimiter=';')
        writer = csv.writer(outfile, delimiter=';')
        for row in reader:
            updated_row = [value.replace(',', '.') if isinstance(value, str) else value for value in row]
            writer.writerow(updated_row)

# Example usage:
input_filename = 'market_analysis_2019.csv'
output_filename = 'market2019.csv'

replace_commas(input_filename, output_filename)
```

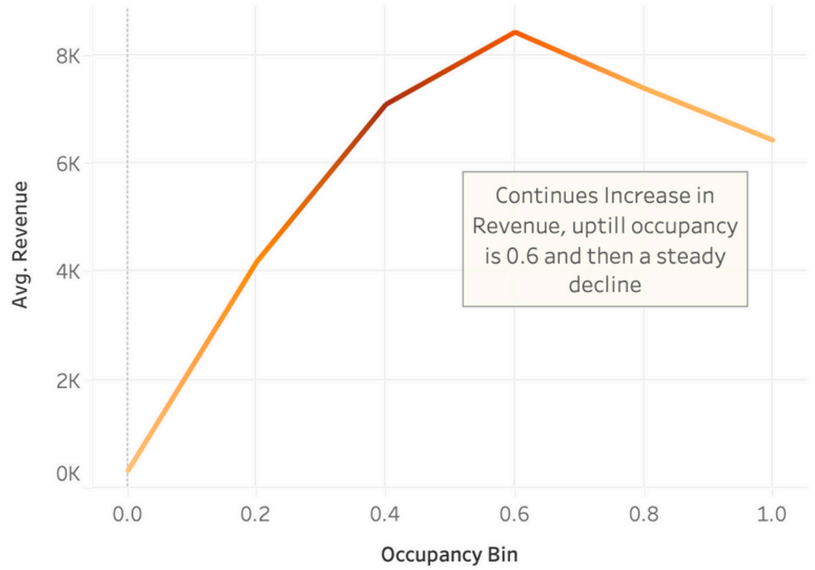
```python
def fill_null_street(input_file, output_file):
    with open(input_file, 'r') as infile, open(output_file, 'w', newline='') as outfile:
        reader = csv.reader(infile, delimiter=';')
        writer = csv.writer(outfile, delimiter=';')
        for row in reader:
            if len(row) > 2 and not row[2]:
                row[2] = "NULL_STREET"
            writer.writerow(row)

# Example usage:
input_filename = 'geo_removedDup.csv'
output_filename = 'geo_removedDup22222.csv'

fill_null_street(input_filename, output_filename)
```

```python
import csv

def remove_duplicates(input_file, output_file):
    seen = set()
    output_rows = []
    with open(input_file, 'r') as infile:
        reader = csv.reader(infile, delimiter=';')
        for row in reader:
            unified_id = row[0]
```

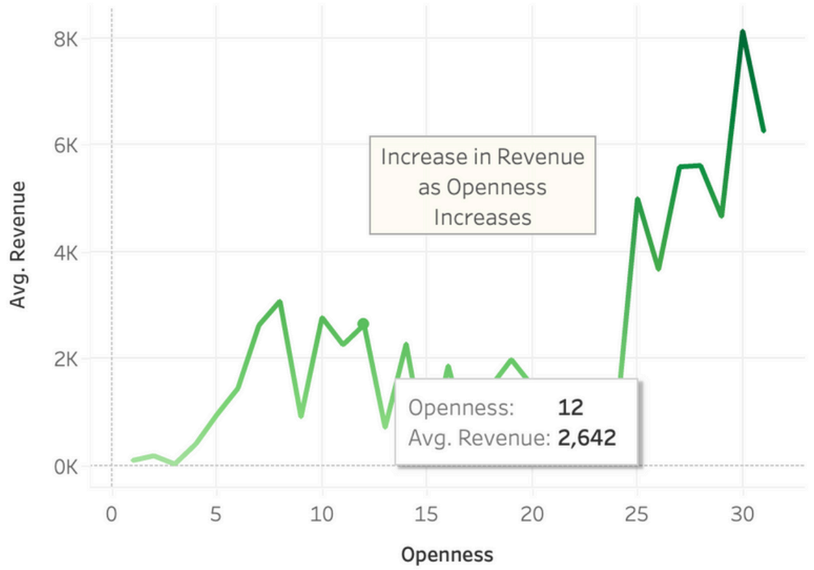
EXPLORATORY ANALYSIS USING TABLEAU

PROFFESIONAL PROPERTIES

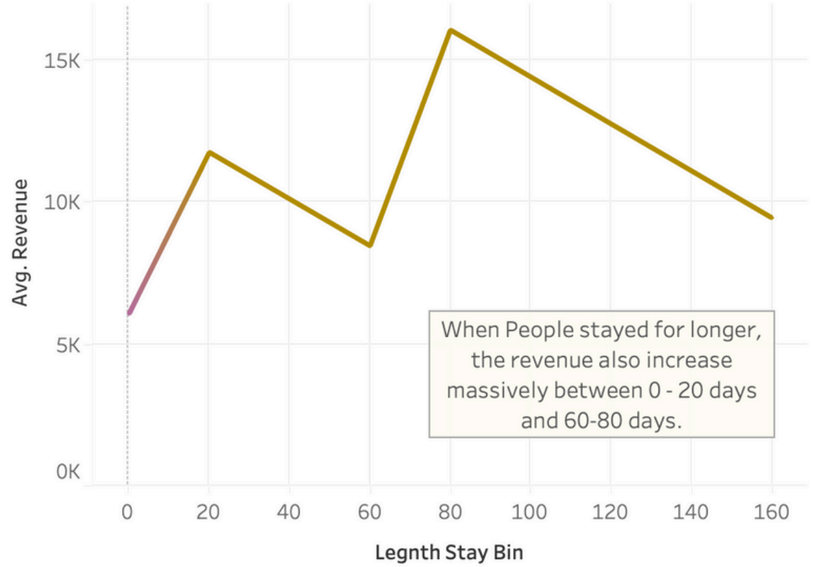
Occupancy vs Revenue



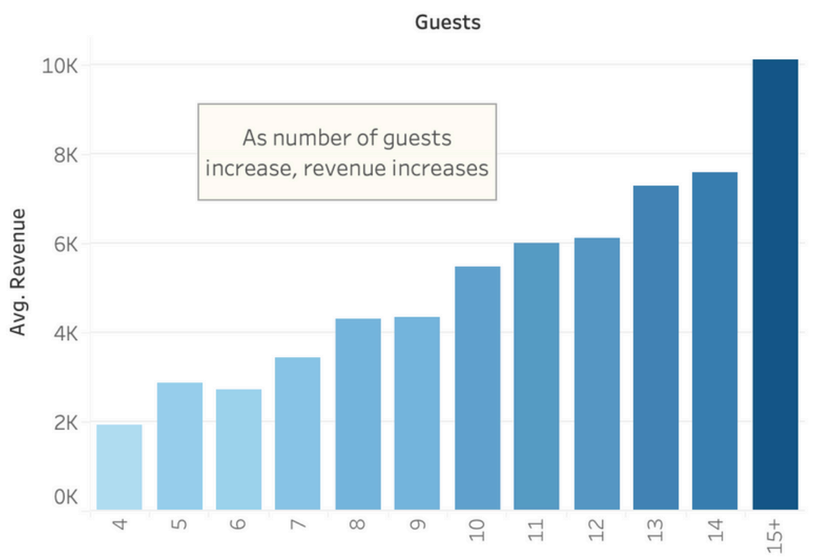
Openness vs Revenue



Length Stay vs Revenue



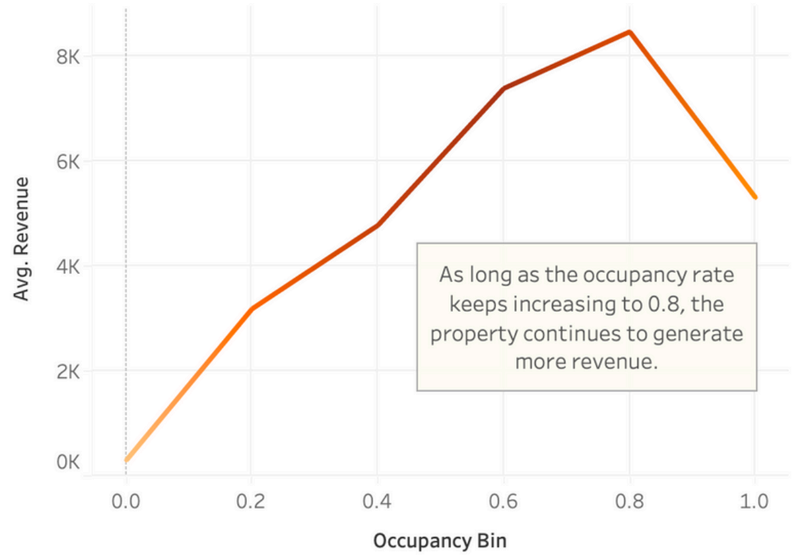
Guests vs Revenue



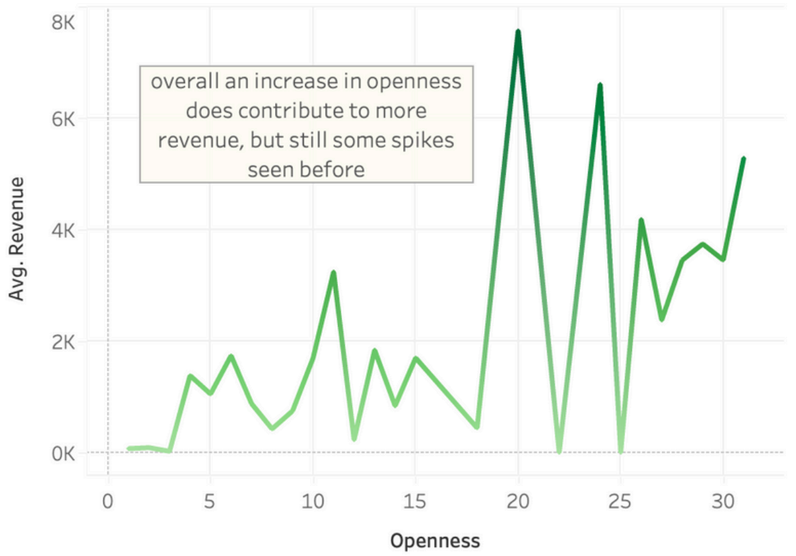
EXPLORATORY ANALYSIS USING TABLEAU

2-5 UNITS PROPERTY

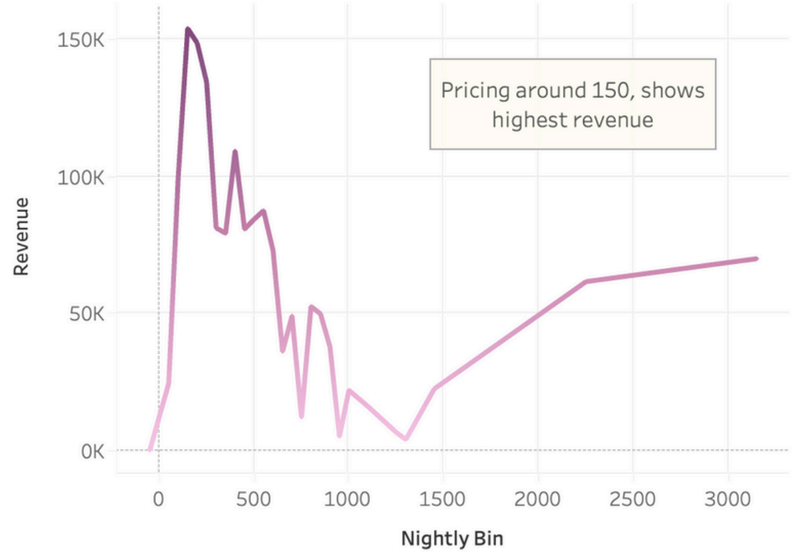
Occupancy vs Revenue



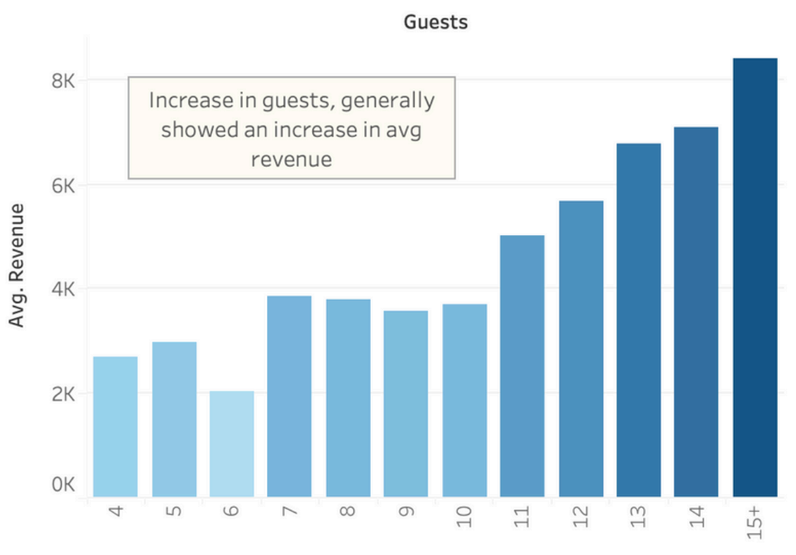
Openness vs Revenue



Nightly Rate vs Revenue



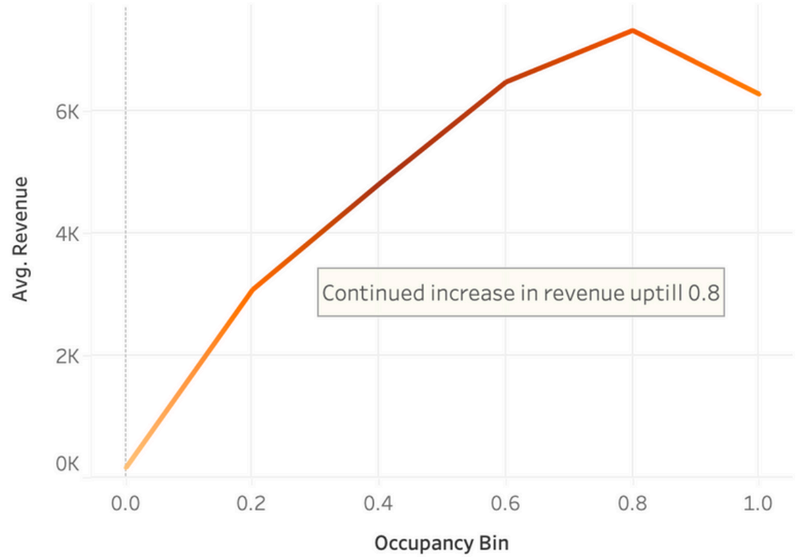
Guests vs Revenue



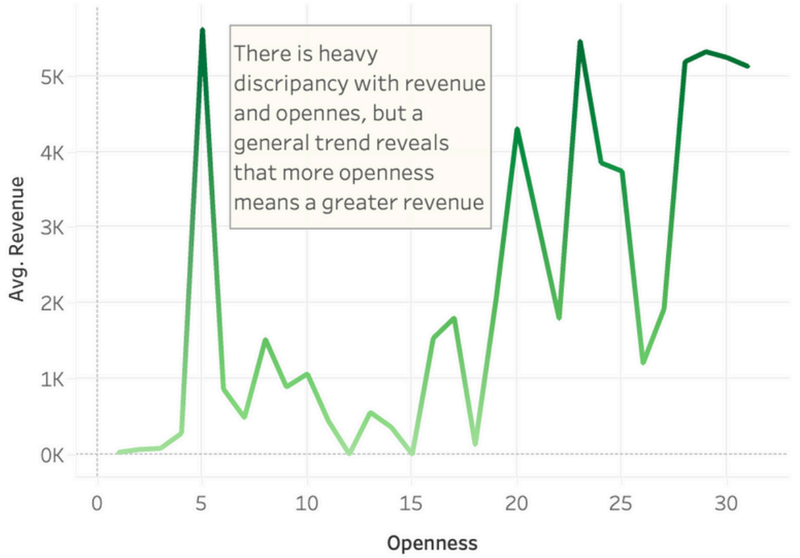
EXPLORATORY ANALYSIS USING TABLEAU

SINGLE OWNERS PROPERTIES

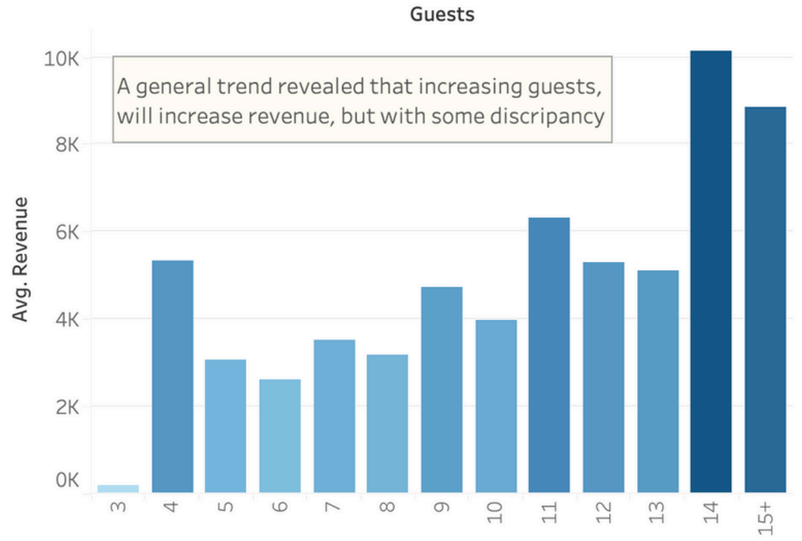
Occupancy vs Revenue



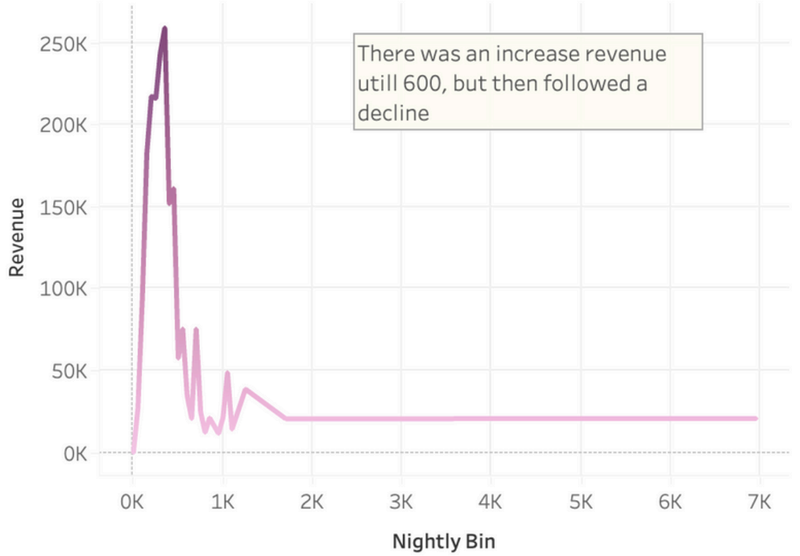
Openness vs Revenue



Guests vs Revenue



Nightly Rate vs Revenue



R STUDIO

MODELING AND ANALYSIS

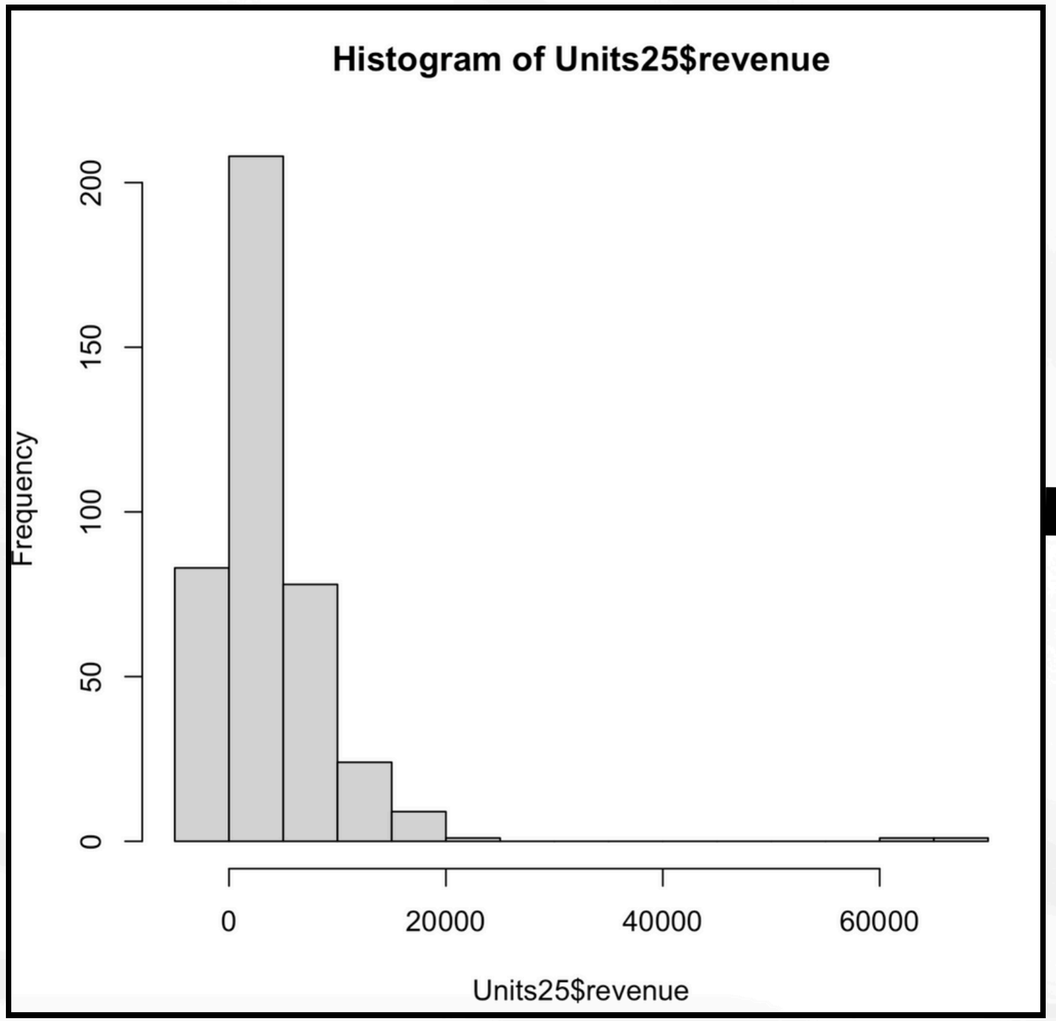
```
'data.frame':   2151 obs. of  20 variables:
 $ unified_id   : chr  "AIR10000347" "AIR10052559" "AIR10178668" "AIR10204420" ...
 $ month        : Date, format: "2019-01-01" "2019-01-01" "2019-01-01" ...
 $ zipcode      : int  92315 92315 92315 92252 92315 92315 92314 92315 92314 92314 ...
 $ city         : chr  "Big Bear Lake" "Big Bear Lake" "Big Bear Lake" "Joshua Tree" ...
 $ bedrooms     : int  3 3 3 3 3 5 3 3 3 5 ...
 $ bathrooms    : num  2 2.5 2.5 2 3 2 1 2 2 4 ...
 $ guests       : num  10 8 7 12 9 12 5 8 8 NA ...
 $ revenue      : num  13949 9947 6027 6549 8610 ...
 $ openness     : int  31 31 31 31 31 31 31 31 31 31 ...
 $ occupancy    : num  1 0.581 0.484 0.452 0.645 ...
 $ nightly.rate : num  450 553 402 468 430 ...
 $ lead.time    : num  8 10.6 7.1 38 43.1 ...
 $ length.stay  : num  65 2.57 1.8 2.43 2.33 ...
 $ Street.Name  : chr  "Cienega Road" "Heavenly Valley Road" "Round Drive" "NULL_STREET" ...
 $ Latitude     : num  34.2 34.2 34.3 34.2 34.2 ...
 $ Longitude    : num  -117 -117 -117 -116 -117 ...
 $ Hot.Tub      : logi  FALSE TRUE FALSE TRUE TRUE FALSE ...
 $ Pool         : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
 $ Professionals: int  0 1 1 0 1 0 0 1 0 0 ...
 $ Single Owners: int  0 0 0 0 0 1 0 0 0 0 ...
```

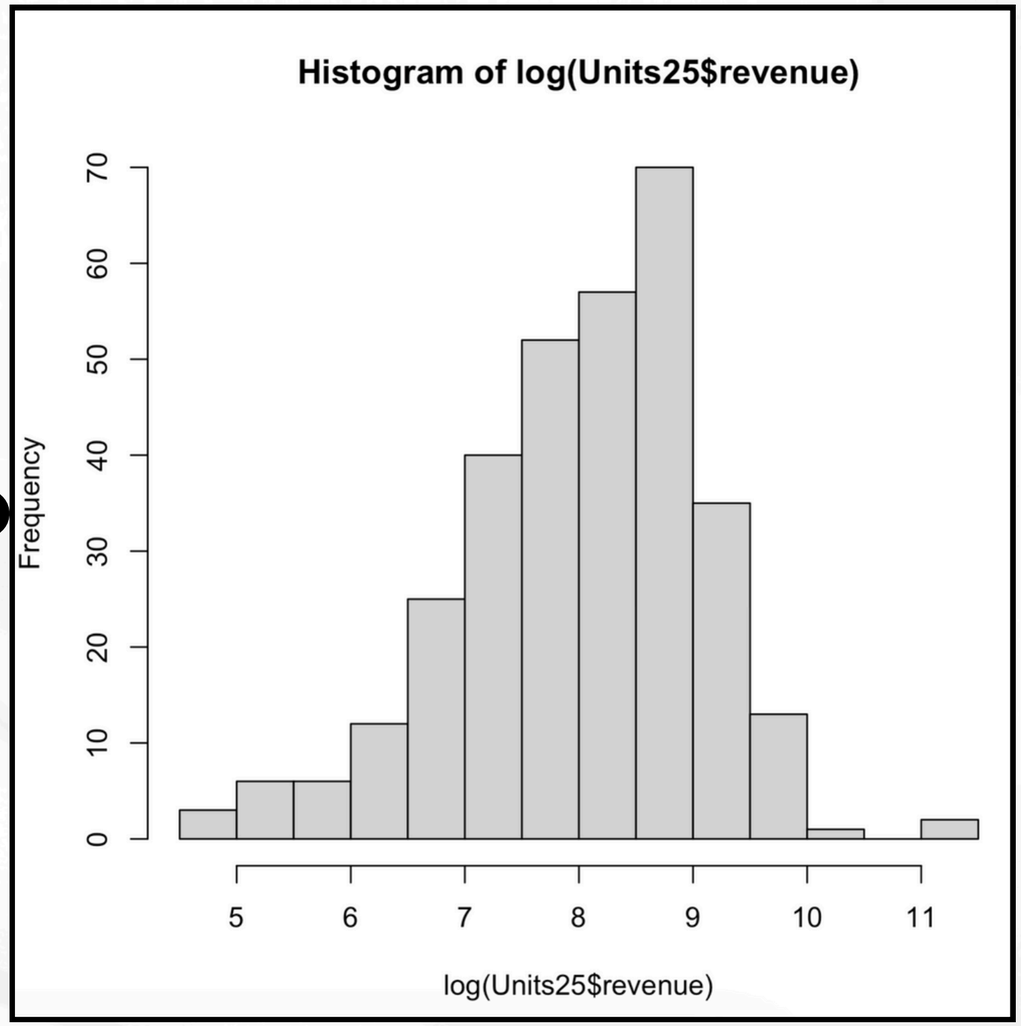
Data Structure

Data explored and used for analysis

R STUDIO

NORMALIZING DATA





PROFESSIONAL PROPERTIES MODEL

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.5766 -0.1084  0.0904  0.2251  0.8034

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.732e+00  1.033e-01  36.118  < 2e-16 ***
bedrooms     -5.870e-02  3.161e-02  -1.857 0.063679 .
bathrooms     3.141e-02  2.518e-02   1.247 0.212617
guests        3.176e-02  8.417e-03   3.773 0.000172 ***
openness      8.058e-02  2.108e-03  38.228  < 2e-16 ***
occupancy     2.522e+00  6.597e-02  38.234  < 2e-16 ***
nightly.rate  2.169e-03  6.541e-05  33.158  < 2e-16 ***
lead.time    -2.098e-03  6.168e-04  -3.401 0.000701 ***
length.stay  -5.281e-04  2.127e-03  -0.248 0.803980
Hot.TubTRUE   6.791e-02  2.866e-02   2.370 0.018019 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4115 on 904 degrees of freedom
  (311 observations deleted due to missingness)
Multiple R-squared:  0.8344,    Adjusted R-squared:  0.8328
F-statistic: 506.2 on 9 and 904 DF,  p-value: < 2.2e-16
```

SINGLE OWNER PROPERTY MODEL

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.42188 -0.23192  0.07169  0.30097  1.09245

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.232e+00  1.841e-01  17.552  < 2e-16 ***
bedrooms      7.947e-02  5.675e-02   1.400   0.1622
bathrooms     1.711e-02  4.468e-02   0.383   0.7020
guests        9.033e-02  1.468e-02   6.155 1.89e-09 ***
openness      8.519e-02  3.555e-03  23.966  < 2e-16 ***
occupancy     2.028e+00  9.733e-02  20.832  < 2e-16 ***
nightly.rate  9.208e-04  6.592e-05  13.968  < 2e-16 ***
lead.time    -1.273e-03  7.074e-04  -1.799   0.0728 .
length.stay  -3.403e-03  2.234e-03  -1.523   0.1285
Hot.TubTRUE   1.321e-01  4.800e-02   2.752   0.0062 **
PoolTRUE     -1.464e-01  1.057e-01  -1.385   0.1668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4622 on 385 degrees of freedom
  (125 observations deleted due to missingness)
Multiple R-squared:  0.7775,    Adjusted R-squared:  0.7717
F-statistic: 134.5 on 10 and 385 DF,  p-value: < 2.2e-16
```

2-5 UNITS MODEL

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.89979 -0.16825  0.07881  0.26516  0.98933

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7997392  0.2152054  17.656  < 2e-16 ***
bedrooms     -0.1136456  0.0625697  -1.816 0.070374 .
bathrooms     0.0369880  0.0516888   0.716 0.474830
guests        0.0507016  0.0143155   3.542 0.000464 ***
openness      0.0798561  0.0037304  21.407  < 2e-16 ***
occupancy     2.0072141  0.1025331  19.576  < 2e-16 ***
nightly.rate  0.0019527  0.0001151  16.967  < 2e-16 ***
lead.time    -0.0013137  0.0007074  -1.857 0.064351 .
length.stay   0.0011432  0.0025975   0.440 0.660205
Hot.TubTRUE   0.2968417  0.0593452   5.002 9.94e-07 ***
PoolTRUE     -0.1257491  0.0898077  -1.400 0.162540
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4597 on 285 degrees of freedom
  (109 observations deleted due to missingness)
Multiple R-squared:  0.8134,    Adjusted R-squared:  0.8069
F-statistic: 124.3 on 10 and 285 DF,  p-value: < 2.2e-16
```

Bank Client and Marketing Outcomes Dataset

Independent Variables

- Age
- Job
- Marital
- Education
- balance
- housing
- duration
- campaign

Dependent Variables

y (Marketing campaign Outcome)

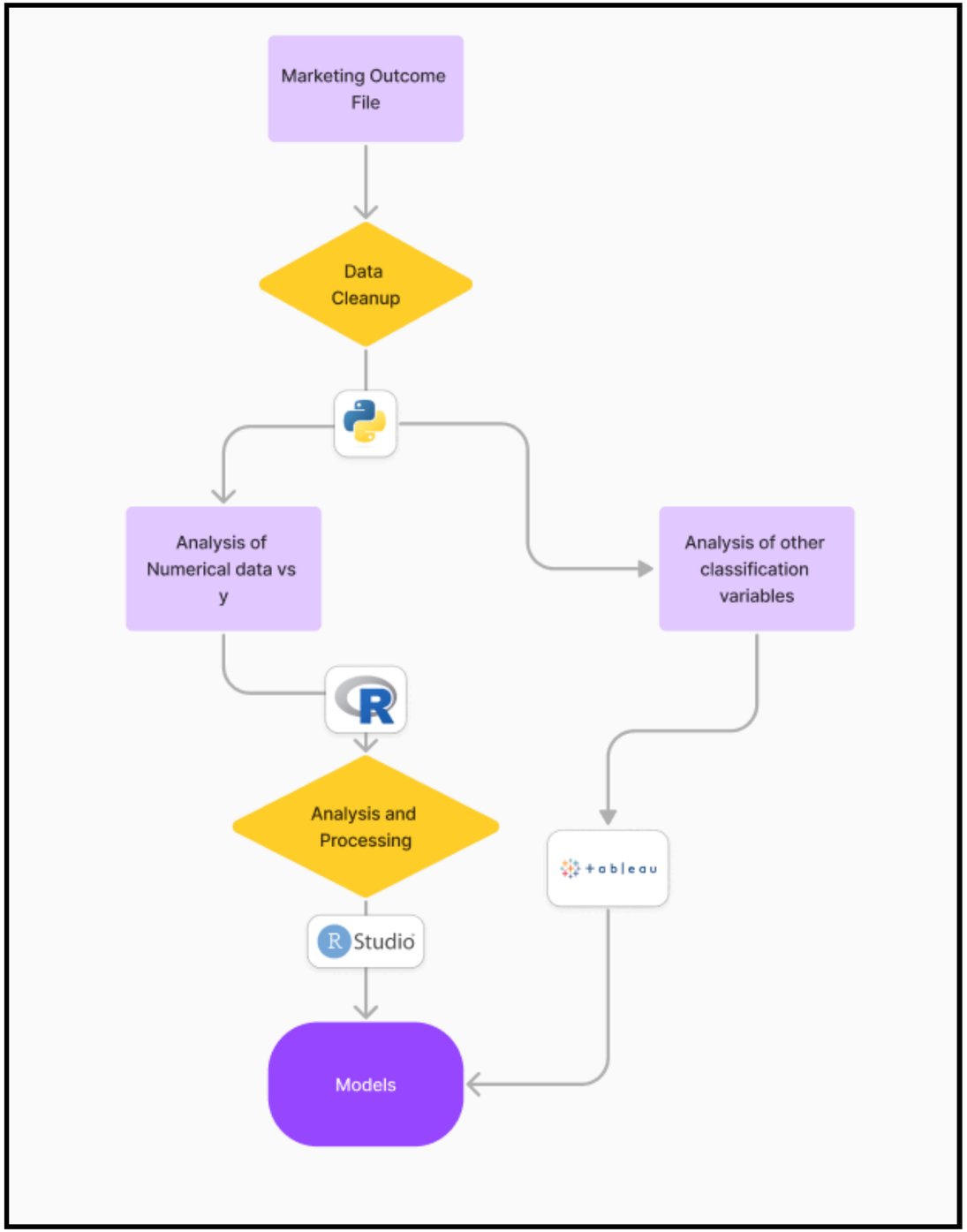
Dataset: https://www.kaggle.com/datasets/computingvictor/zillow-market-analysis-and-real-estate-sales-data

INTRODUCTION

PROCESS AND APPROACH

Tools Used:

Tableau -> Data analysis and visualisation
R -> Analysis and Logistic Regression Model
Python -> Data Cleaning



INTRODUCTION

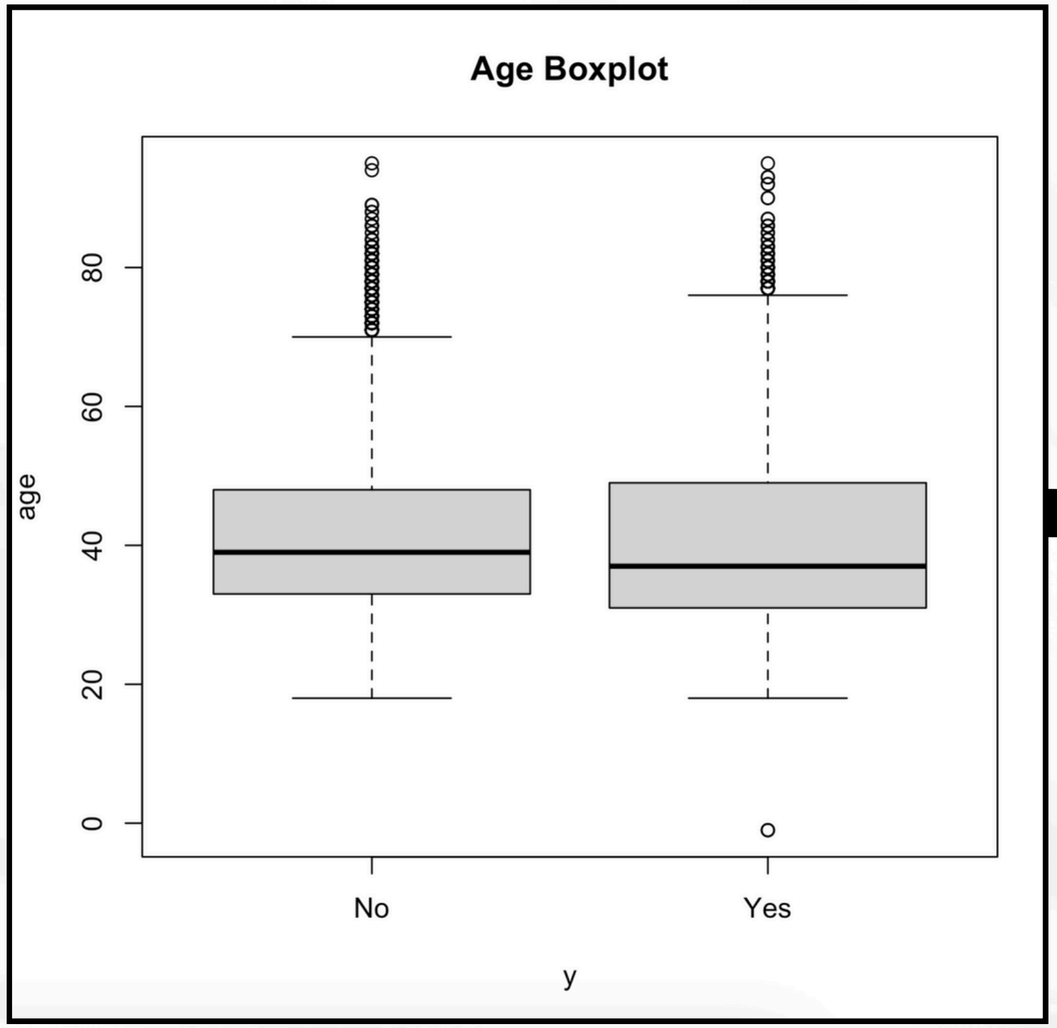
DATA CLEANING

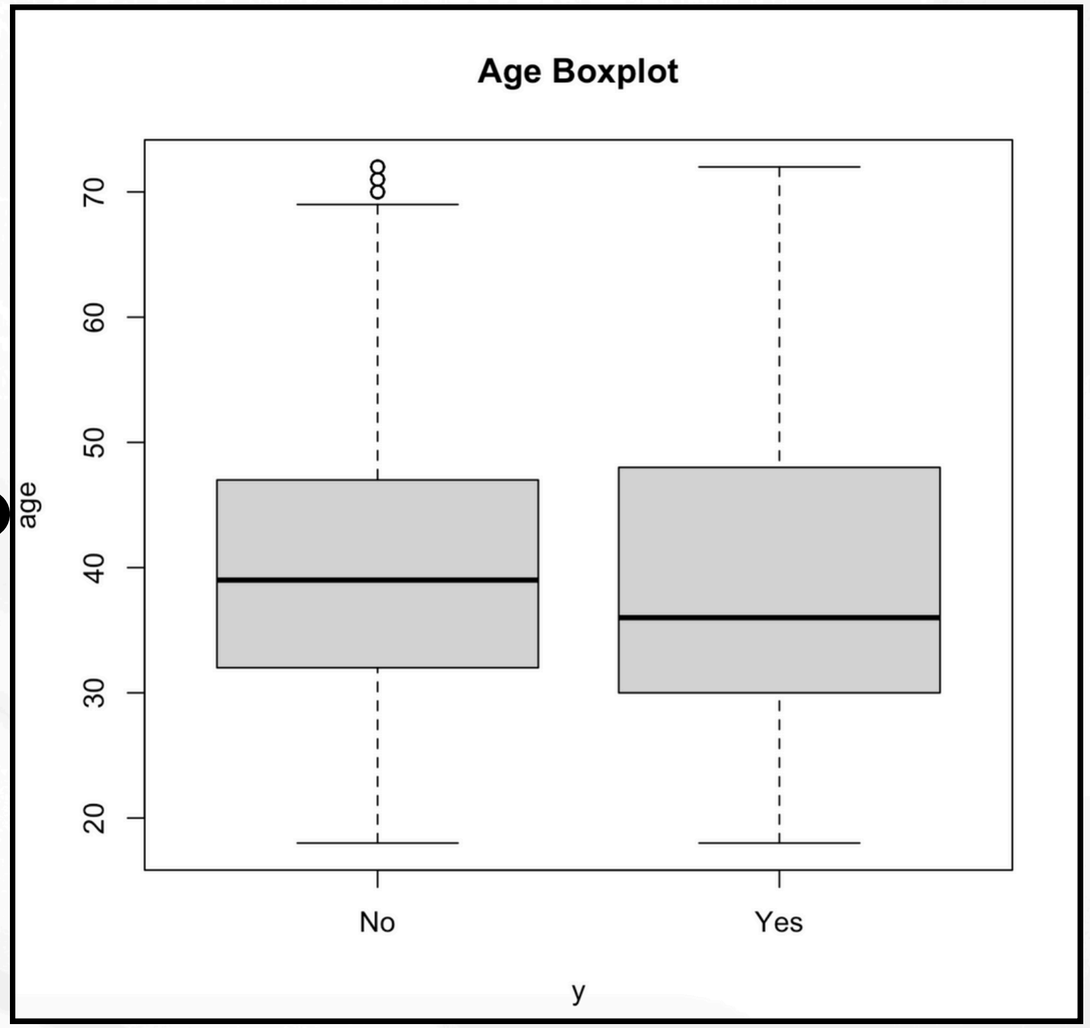
1. Removing Null data values

```python
import pandas as pd

def remove_null_rows(input_file, output_file):
    # Read CSV file
    df = pd.read_csv(input_file)

    # Remove rows with null values
    df = df.dropna()

    # Write cleaned data to a new CSV file
    df.to_csv(output_file, index=False)

# Example usage
input_file = 'dataset.csv'
output_file = 'output.csv'
remove_null_rows(input_file, output_file)
```

R STUDIO - DATA CLEANING

REMOVING OUTLIERS





SUMMARY OF GLM MODELS

Variable	Strength of Correlation	Pr(>|z|)	AIC
balance	99.99	<2e-16	20370
duration	99.99	<2e-16	18493
day	99.99	4.16e-14	20589
age	90	0.0999	20644

FINAL MODEL

```
Call:
glm(formula = y ~ balance + duration + day + age, family = binomial,
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.771e+00  1.041e-01 -36.225  < 2e-16 ***
balance      4.590e-04  3.078e-05  14.913  < 2e-16 ***
duration     5.684e-03  1.294e-04  43.921  < 2e-16 ***
day         -1.396e-02  2.554e-03  -5.468 4.56e-08 ***
age         -5.310e-03  2.111e-03  -2.515   0.0119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18806  on 34994  degrees of freedom
Residual deviance: 16593  on 34990  degrees of freedom
AIC: 16603

Number of Fisher Scoring iterations: 6
```

TESTING THE MODEL

```
   The Hosmer-Lemeshow goodness-of-fit test

 Group Size Observed  Expected
     1 3500        8  64.05068
     2 3500       34  85.94777
     3 3500       78 104.94351
     4 3500      116 125.56851
     5 3500      152 150.14225
     6 3500      218 183.49683
     7 3500      286 230.35184
     8 3500      373 306.16127
     9 3500      558 460.95133
    10 3495      834 945.38601

          Statistic =  168.8133
degrees of freedom =  8
            p-value =  < 2.22e-16
```

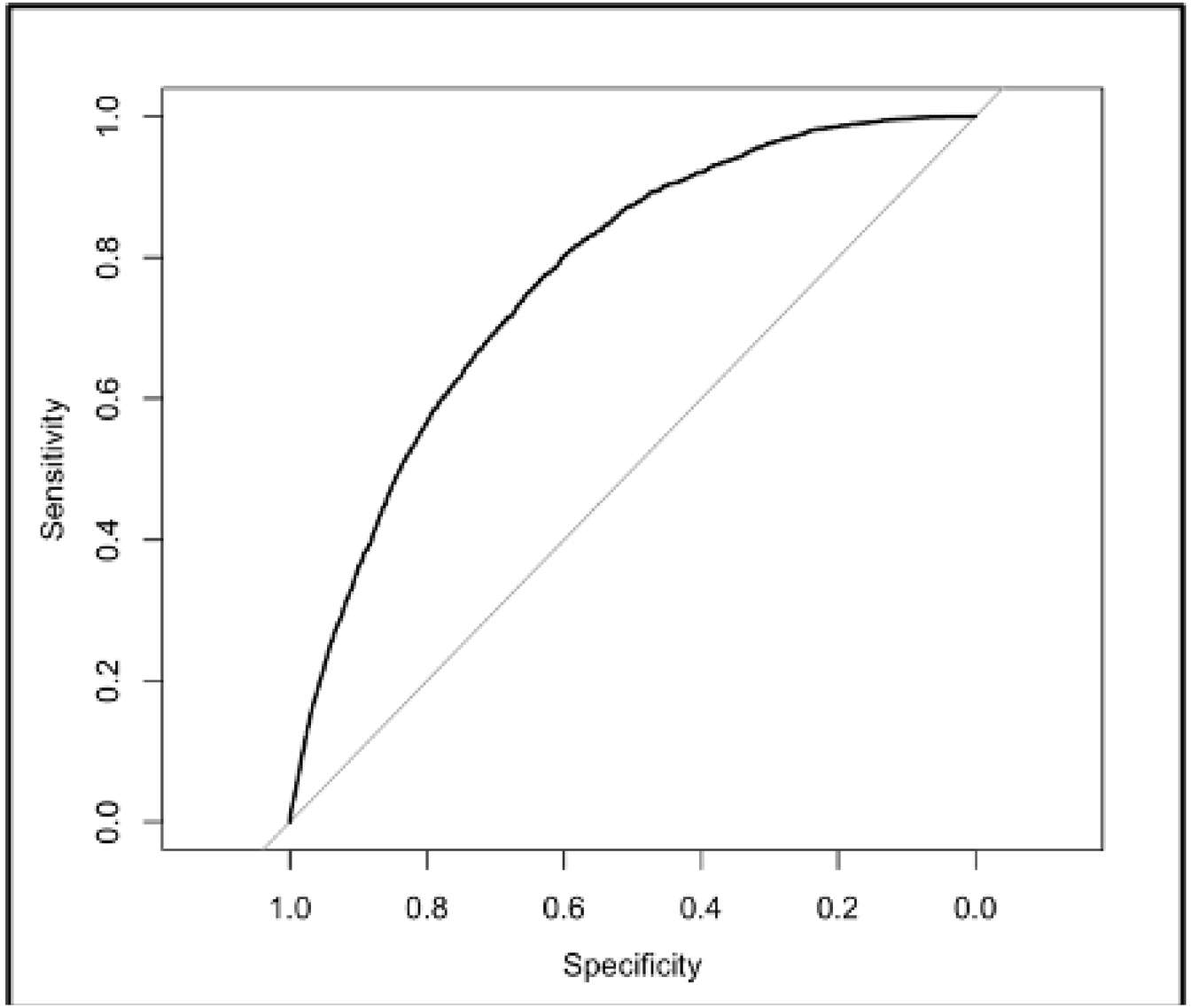
CONFUSION MATRIX

	Actual Values	
Predicted Values	**No**	**Yes**
No	**9887**	**94**
Yes	**58**	**38**

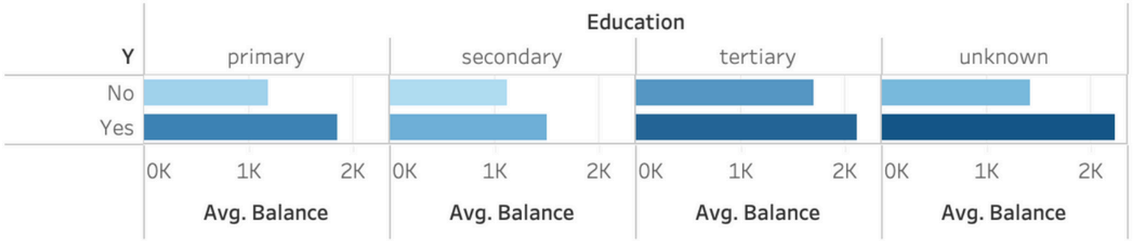
False Positive: 0.00583

Total Error Rate: 0.0152

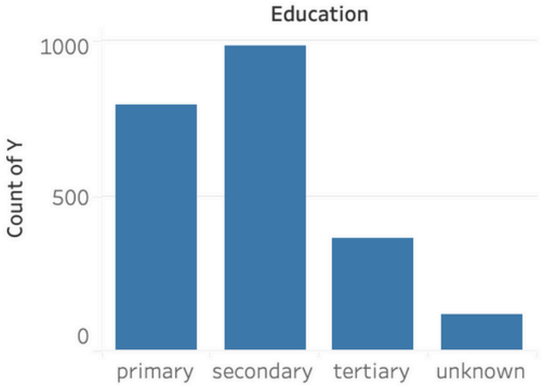
ROC CURVE



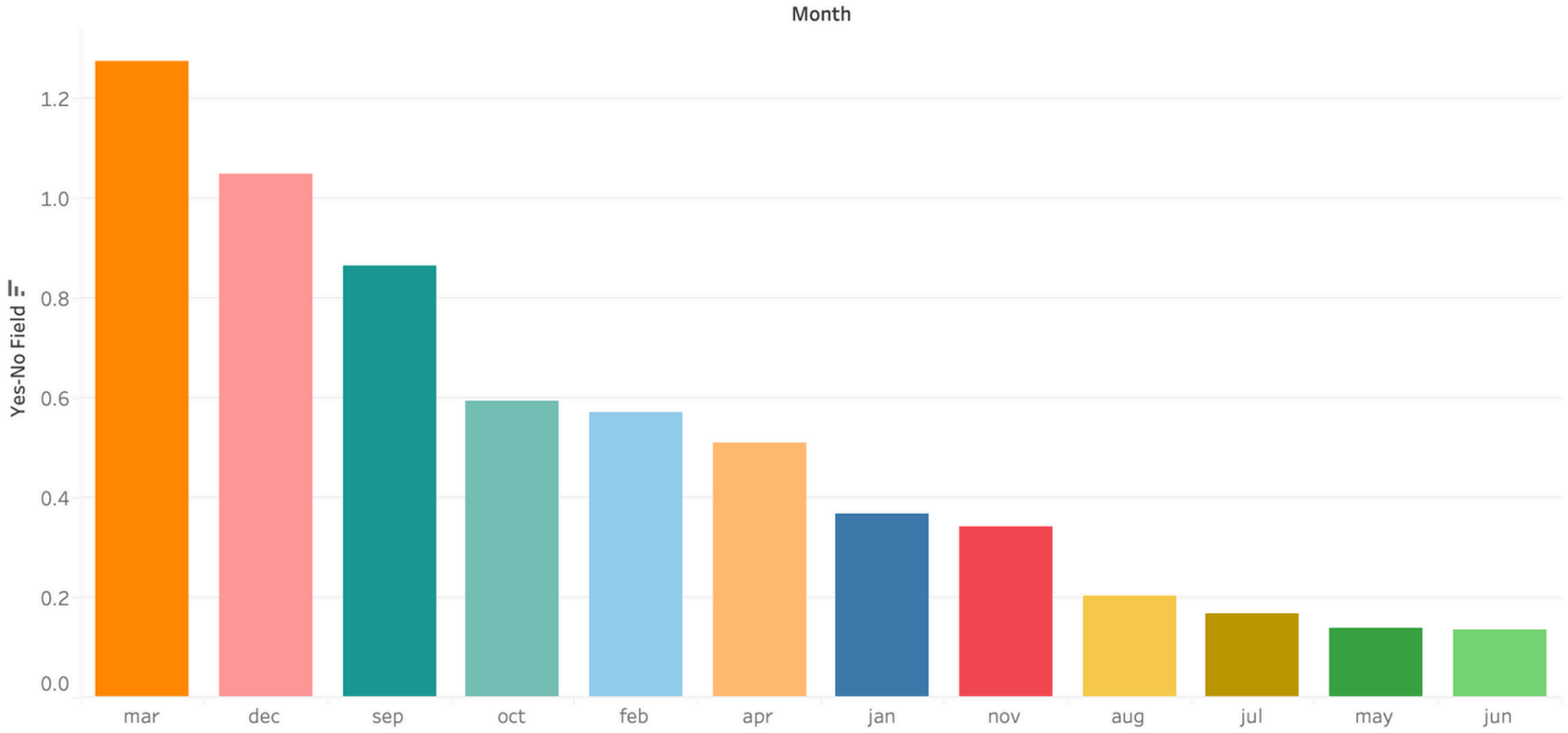
Balance and Education



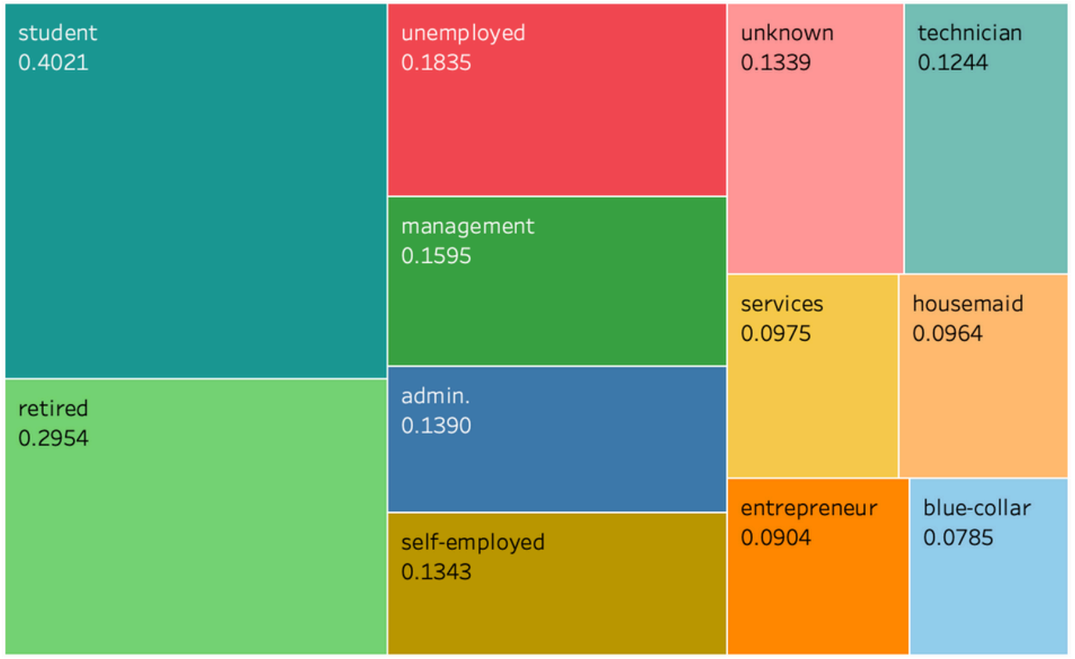
Acceptance Count



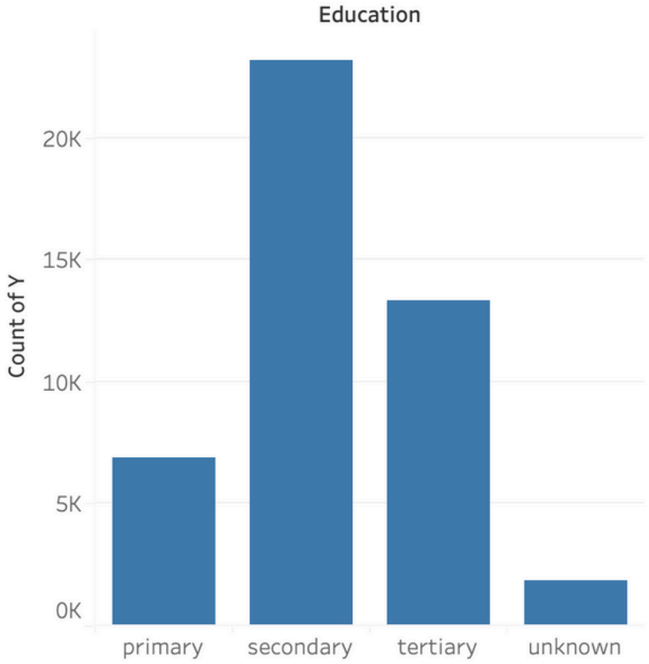
Month vs Acceptance rate



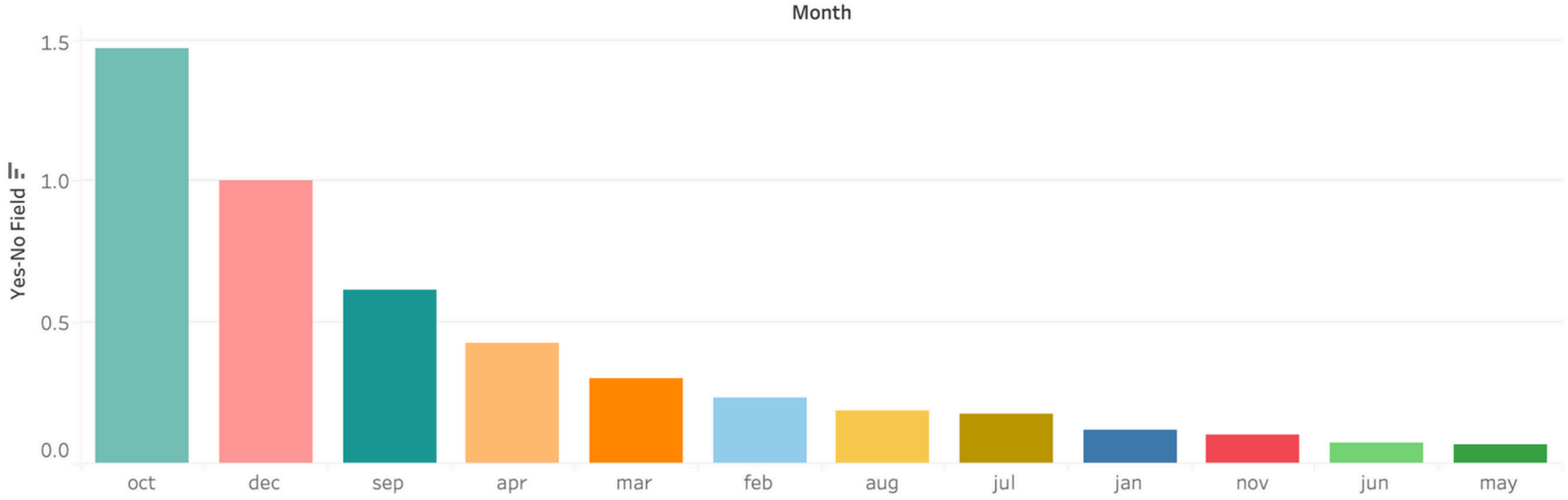
Job against Acceptance



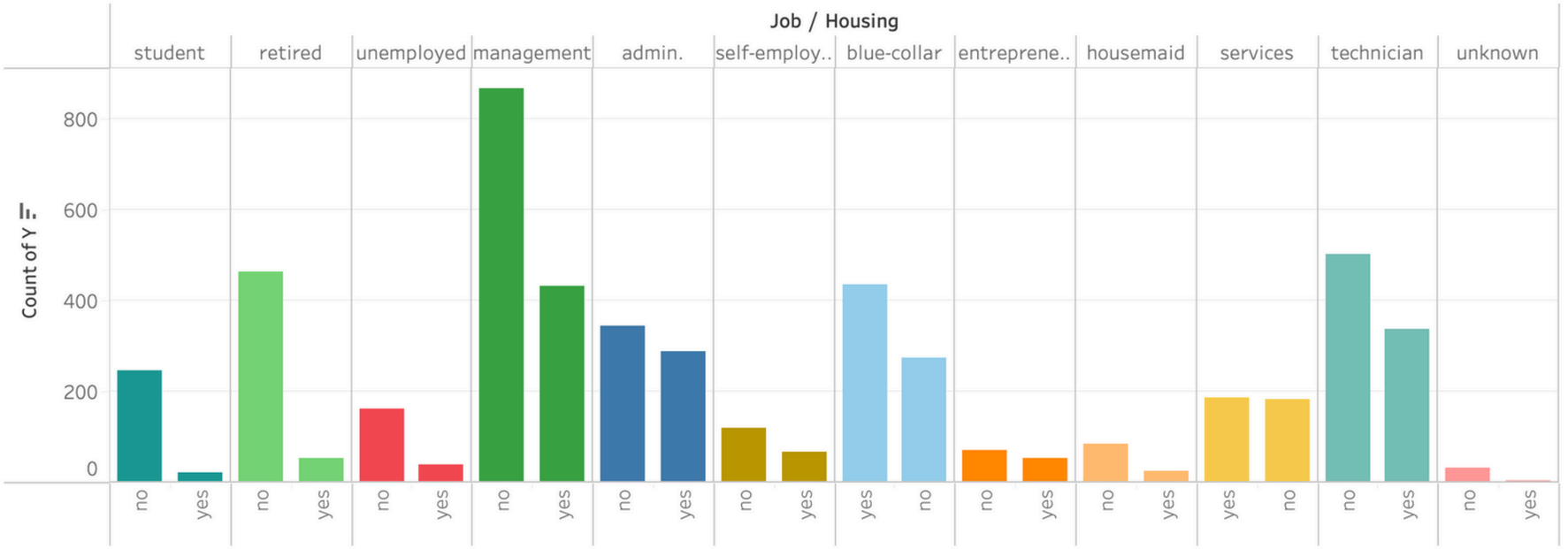
Acceptance Count



Month vs Acceptance rate



Housing for top Job accepters



Marital Status for Job Type