Project Portfolio Milestone

- Applied Data Science
- School of information studies
- Zeyang Zhou

Agenda

- Introduction
- IST 659 – Data Administration and Database Management
- IST 687 – Introduction to Data Science
- IST 652 – Scripting for Data Analysis
- IST 707 – Data Analytics
- Conclusion

Introduction

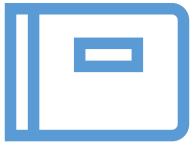
- Self introduction
- Program description
- Learning experience

IST 659 – Data Administration and Database Management

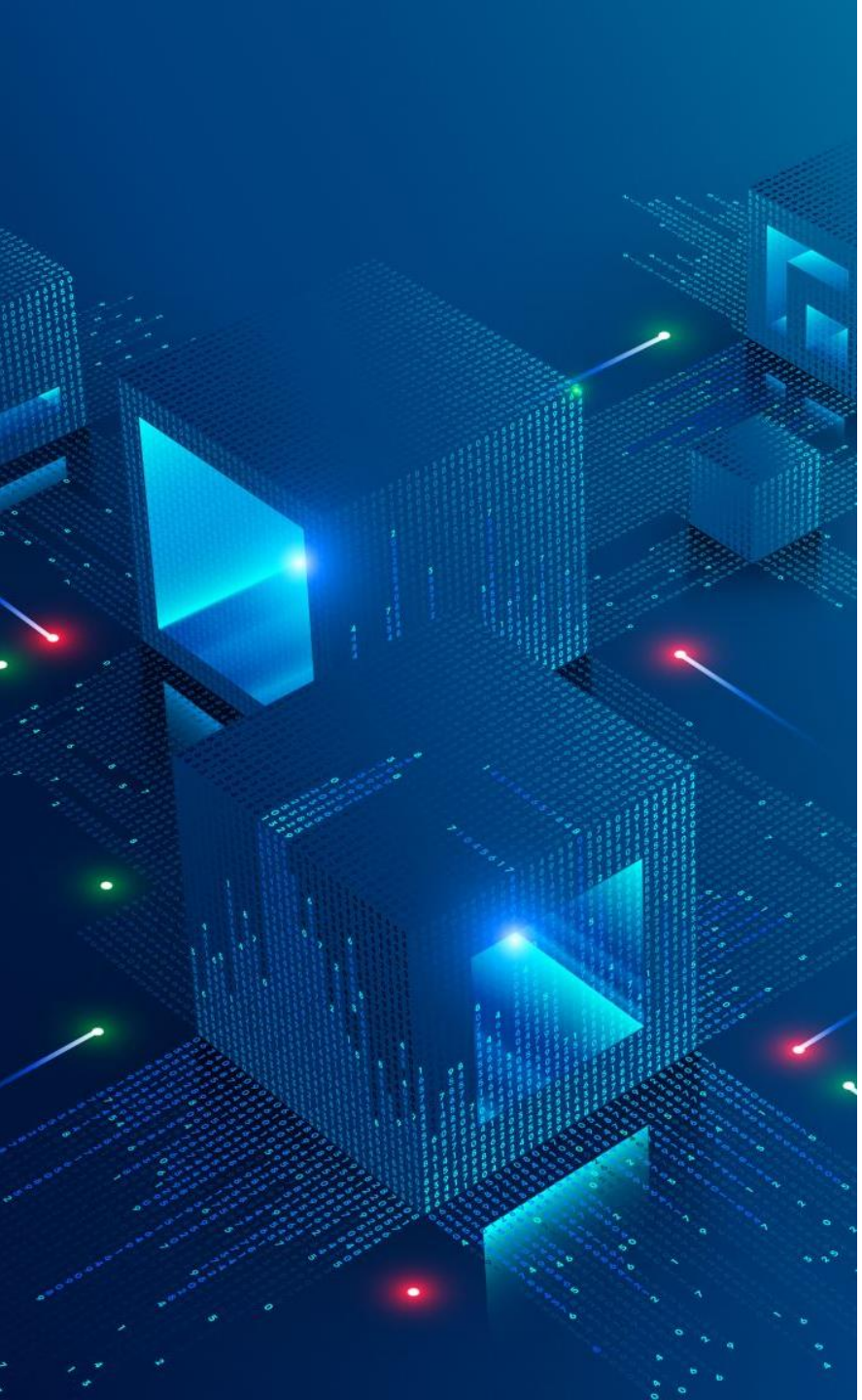
Project overview

Domain knowledge

How to design a database and a matched front end



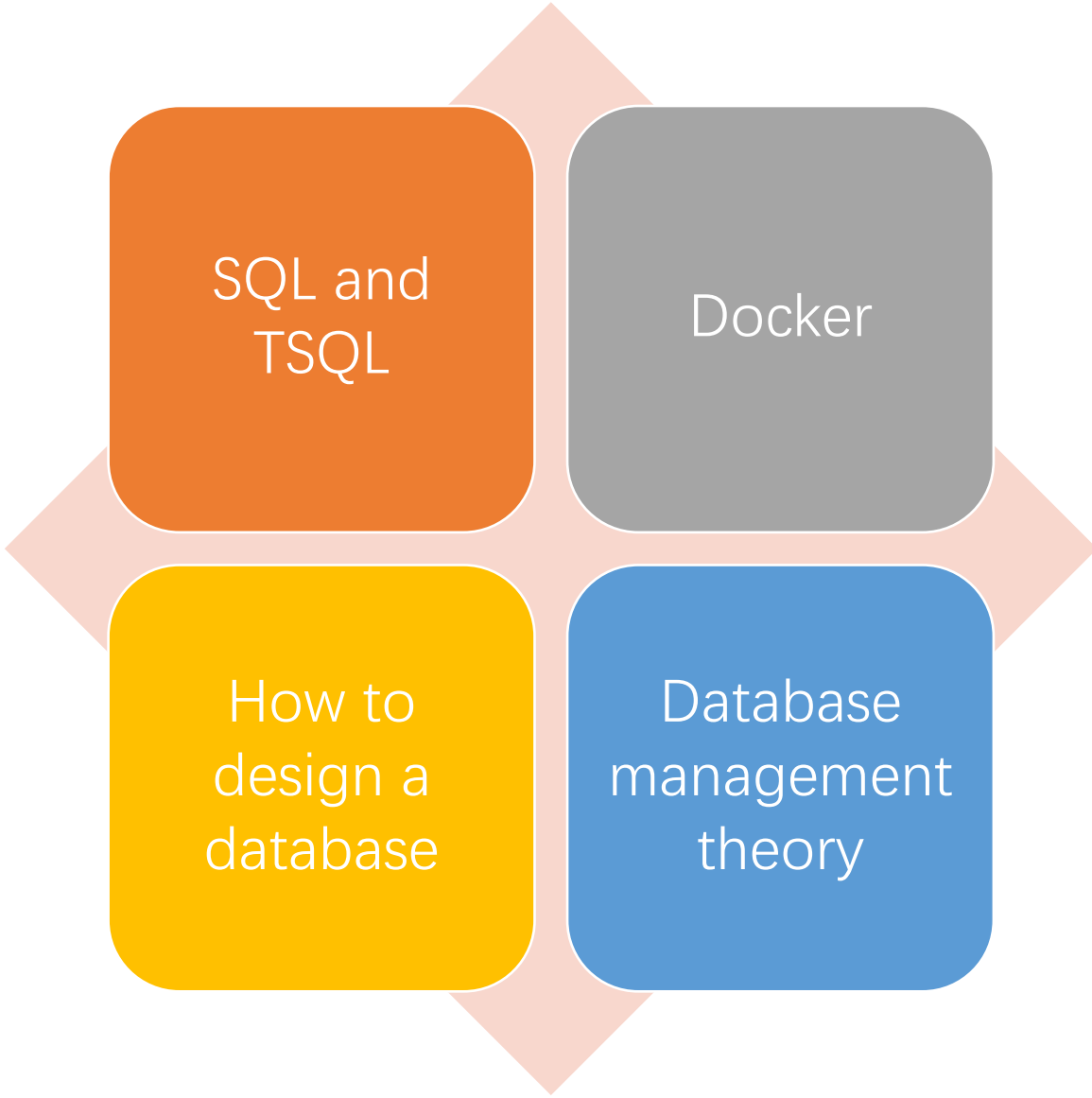
Demo



Project overview

- Azure Data Studio
- Covid-19
- Information system
- infectious diseases

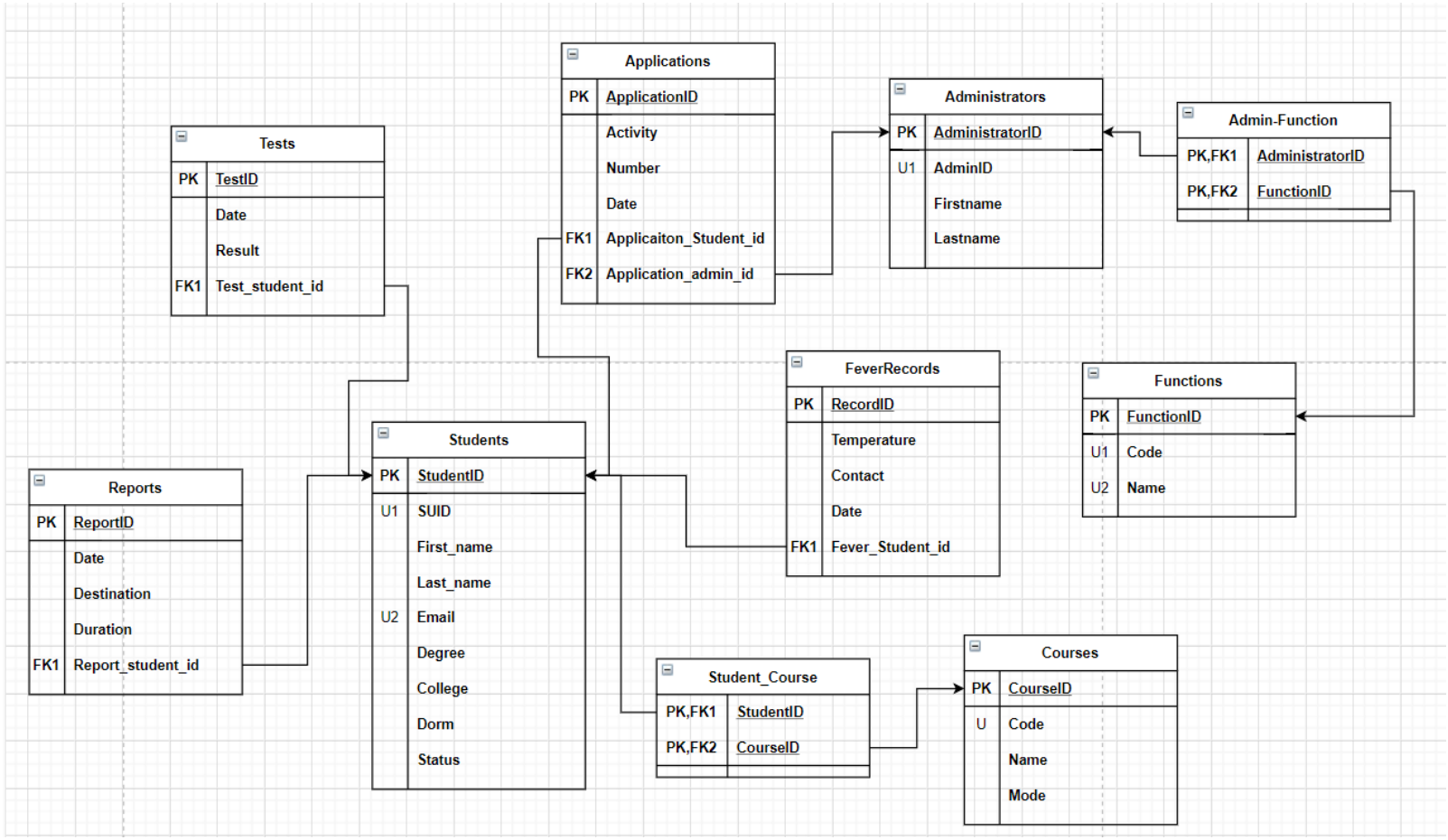
Domain Knowledge



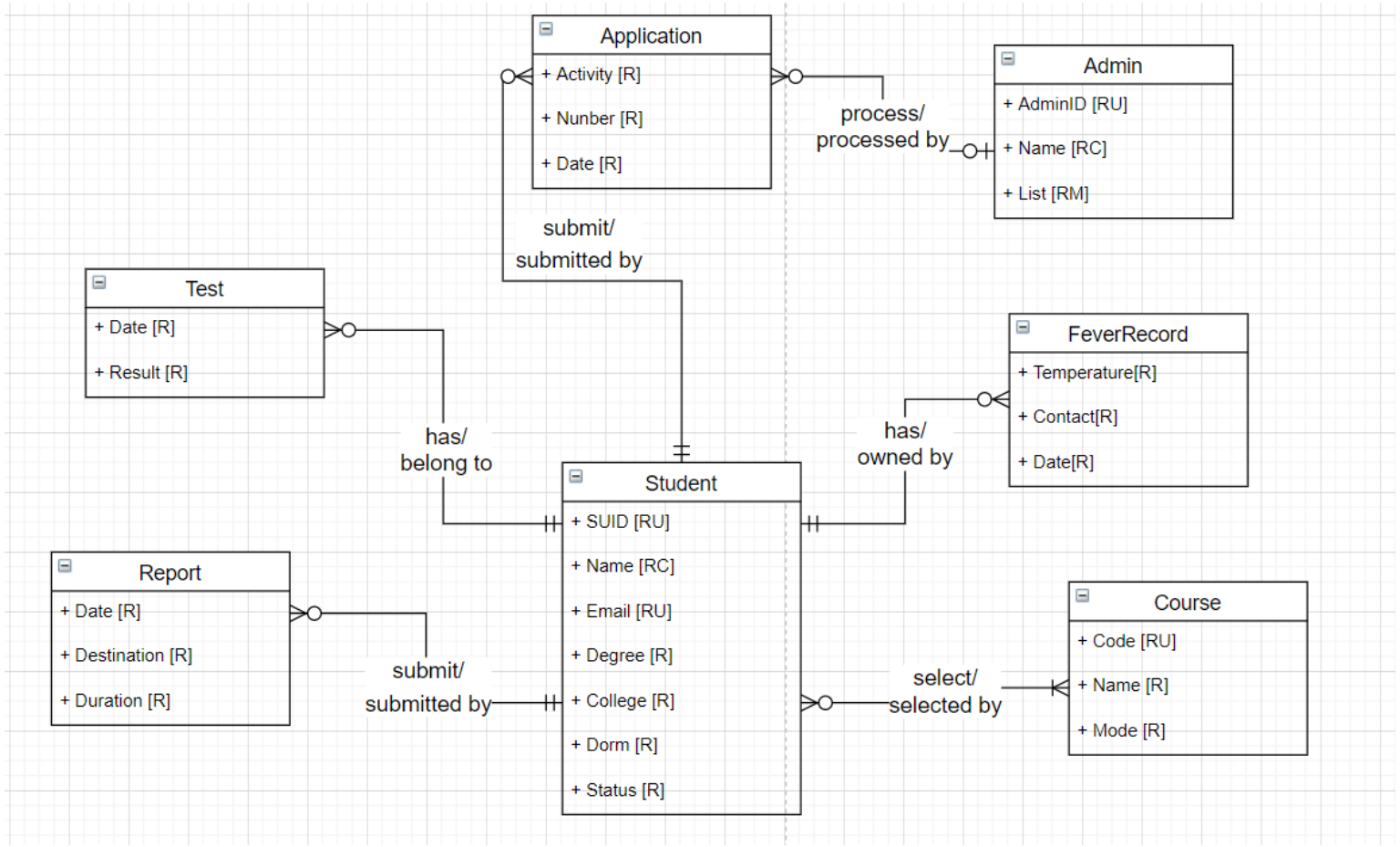
ERD

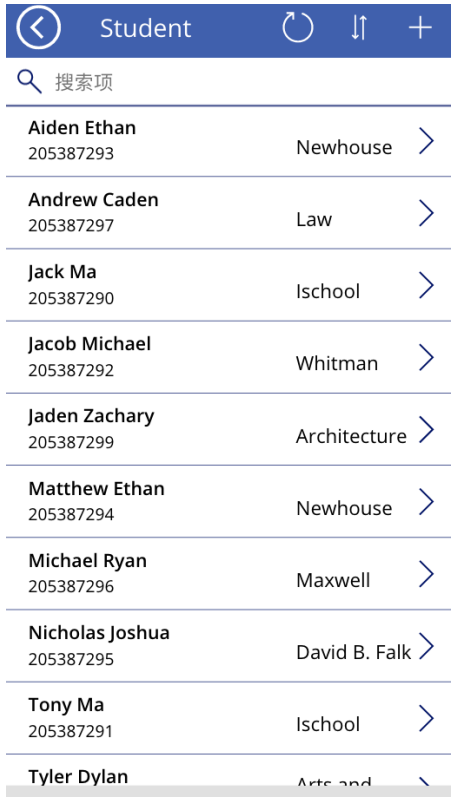
	College	R	Ischool, Whitman, Health etc.		**report**	submitted by	1	1	**student**
	Dorm	R	living on campus or off cumpus						
	Status	R	Red: positive and Green: negative	student-application	**student**	submit	0	M	**application**
					application	submitted by	1	1	**student**
Test	Date	R	Test date						
	Result	R	positive or negative	student-fever	**student**	has	0	M	**feverrecord**
					feverrecord	owned by	1	1	student
Report	Date	R	date of departure						
	Destination	R	Los Angeles, Washington D.C. etc.	student-course	**student**	select	1	M	**course**
	Duration	R	How long, e.g. 3,4 ,5 days		**course**	selected by	0	M	**student**
Application	Activity	R	More than 5 people	Admin-Application	**admin**	process	0	M	**application**
	Number	R	The number of gatherings		**application**	processed by	1	1	**admin**
	Date	R	Time						
FeverRecord	Date	R	The time of onset of fever						
	Contact	R	Gatherings avtivities or not. Yes or No						
	Temparature	R	Fahrenheit degree, >100 degrees						
Course	Coursecode	RU	e.g. SCM651, IST659						
	Coursename	R	e.g. Database, data analysis						
	Mode	R	In person or online						
Administrator	AdminID	RU	similar to SUID						
	Name	RC	Frist name and last name						
	Function	RM	power list						

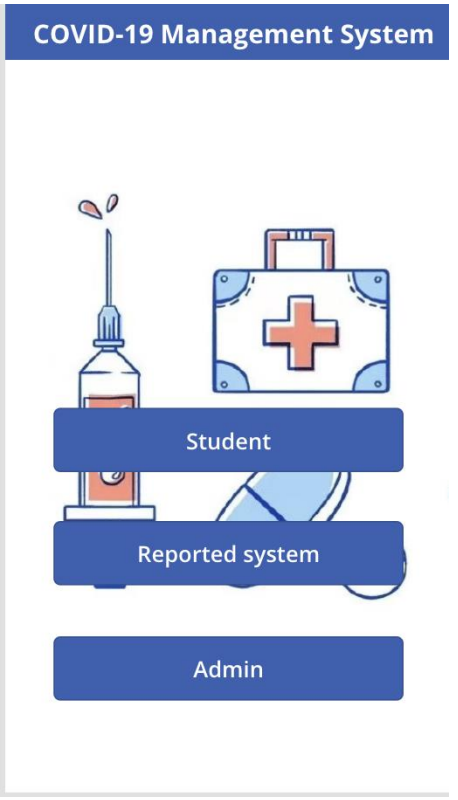
Conceptual Model



Logical Model







UI design and Demo

Course Status

搜索项

lst 659
Online

lst 615
In person

lst 618
Online

lst 422
Online

lst 959
Online

lst 678
In person

Test Report

搜索项

Jack Ma
2020/9/10
negative

Jack Ma
2020/9/10
negative

Tony Ma
2020/9/11
negative

Jacob Michael
2020/10/11
negative

Jack Ma
2020/10/11
negative

Make Travel Report

* Travel Location:

* Reported By Student
Aiden Ethan

* Reported Date
2001/12/31

* Duration
50

Travel Request

Reported By Student:
Aiden Ethan

Travel Location:
Augusta

Reported Day:
6

Reported Date
2020/10/11

Pending Activity

搜索项

Jack Ma
6
Party

Jack Ma
6
Party

Jacob Michael
10
Party

Aiden Ethan
13

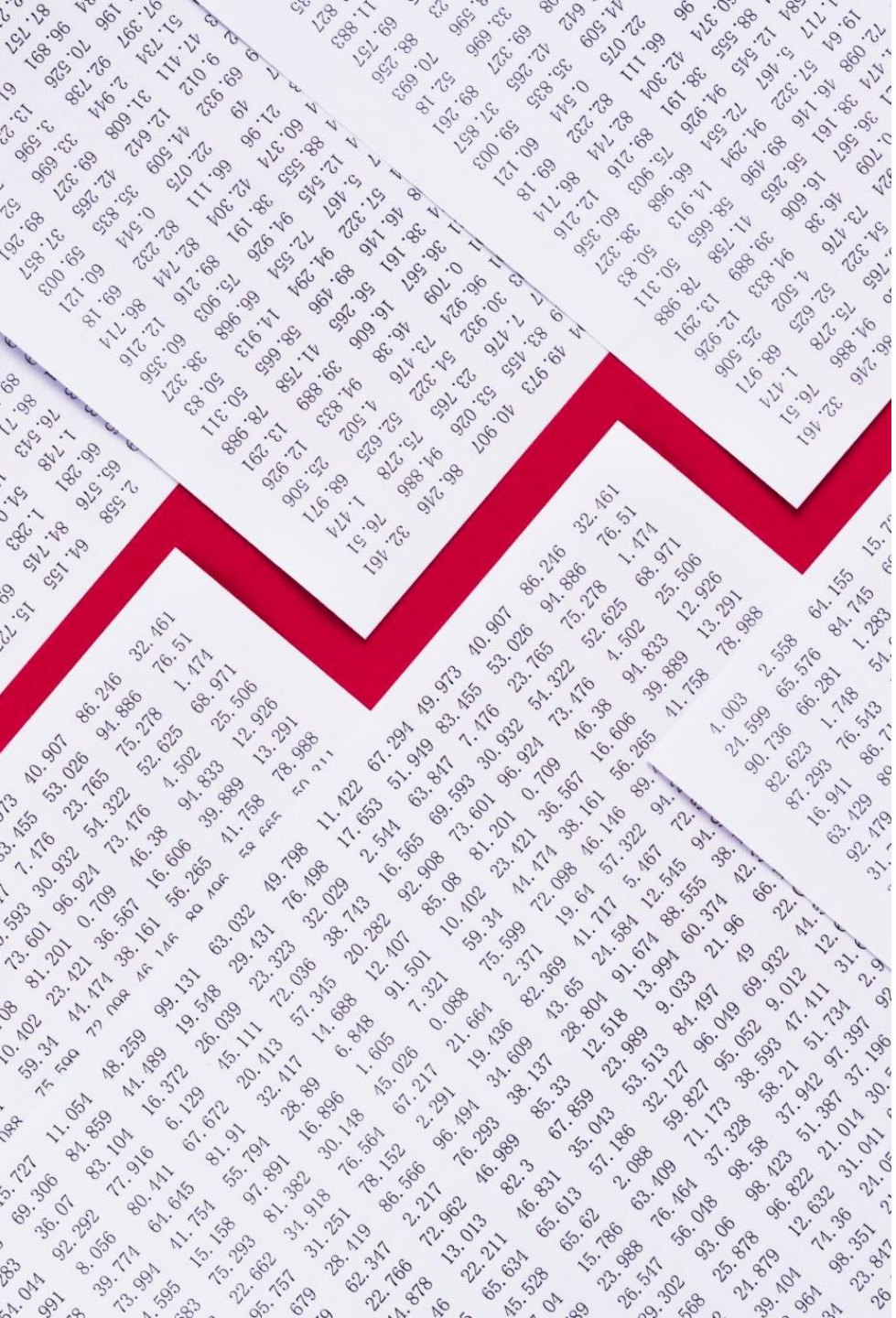
UI design and Demo

IST 687 – Introduction to Data Science

- Project overview
- Customer Churn in the airline Industry
- Visualization and modeling techniques
- Overall Interpretation of Results

Project overview

- Southeast Airlines needed to lower their customer churn (sometimes referred to as customer attrition).
- R and Rstudio
- Introduction to data analysis
- Machine learning
- Report

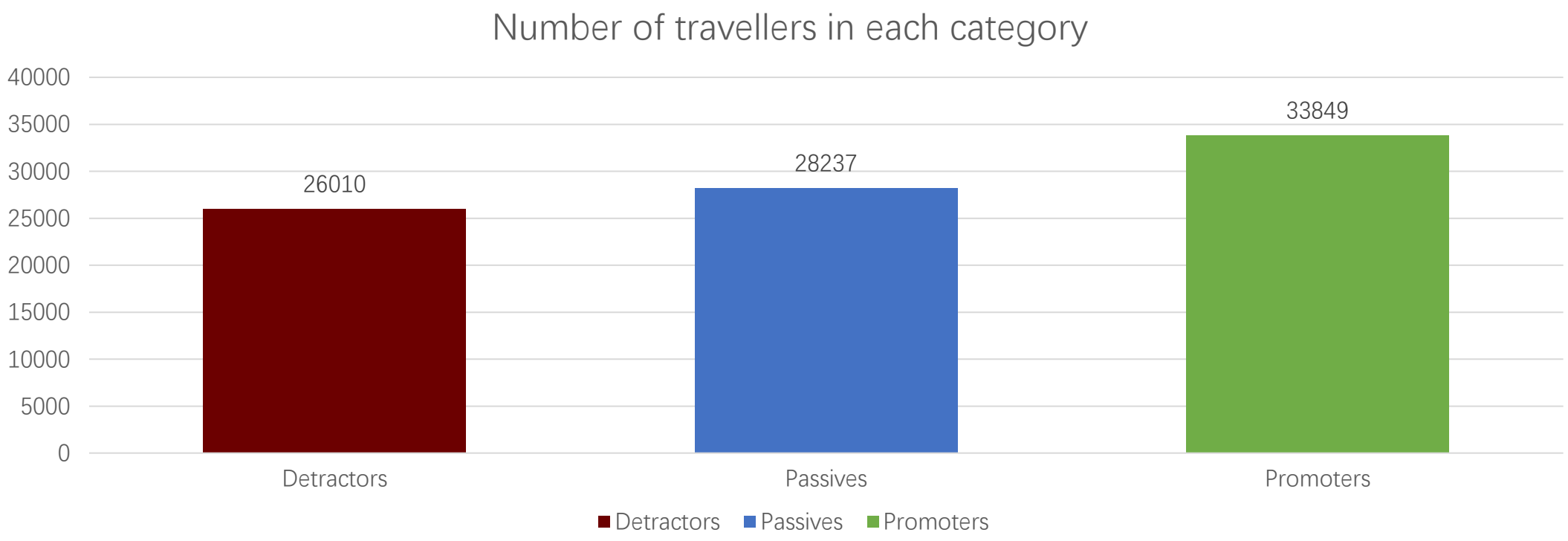


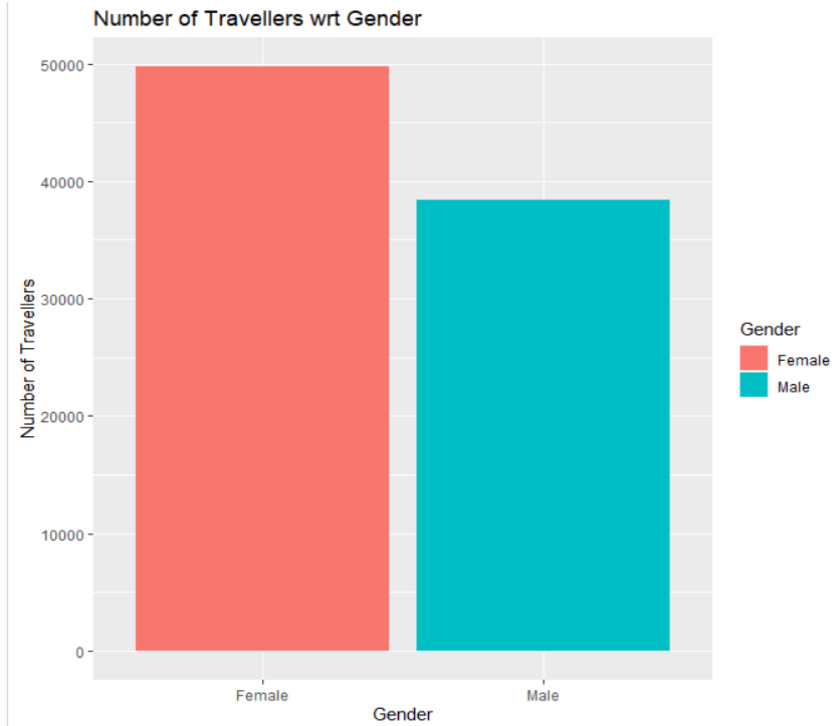
Customer Churn

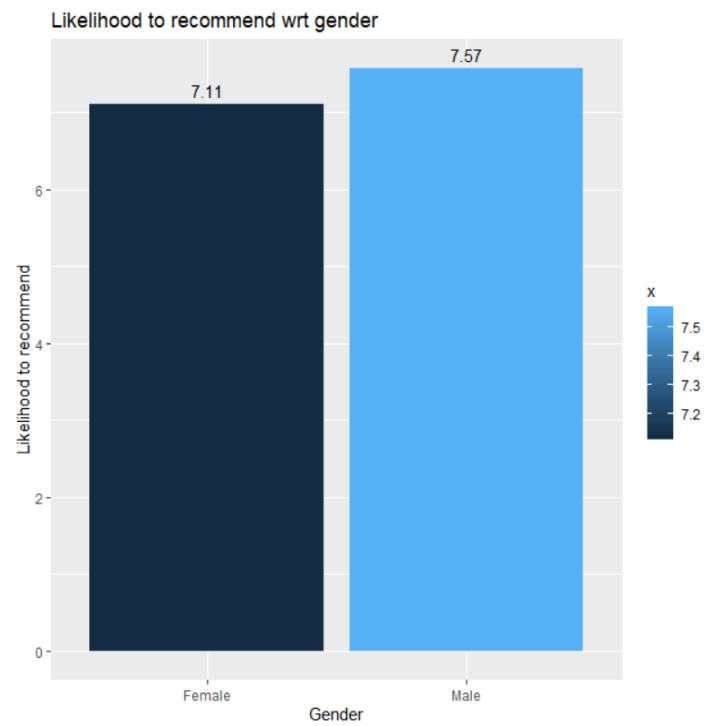
- Customer churn is actually a lagging indicator, meaning the loss has already occurred. As such, it was a measurement of the damage inflicted. The real goal is to reduce churn by getting ahead of the loss (of the customer) by identifying some leading indicators, or metrics, that might help keep a customer.

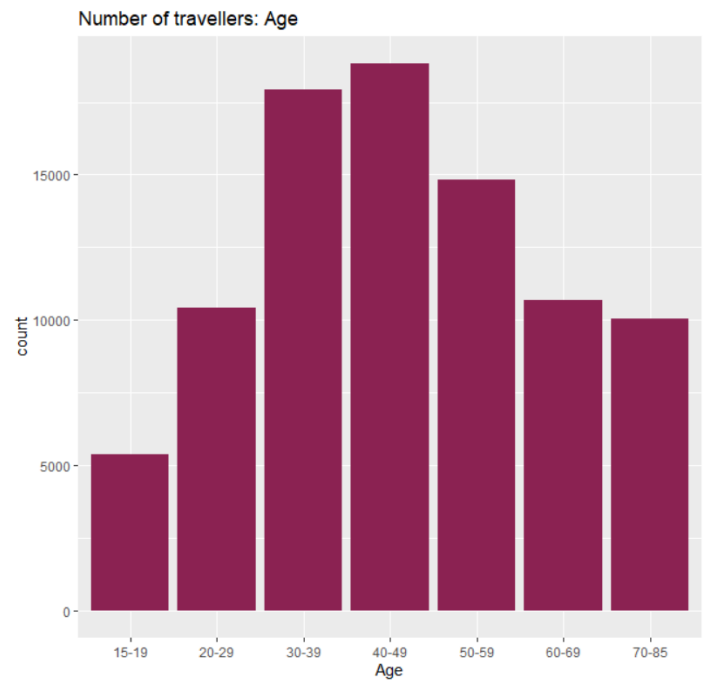
- Indicators:
- Net Promotor Score (NPS), Southeast and its Regional Airline Partners, The Data Available…

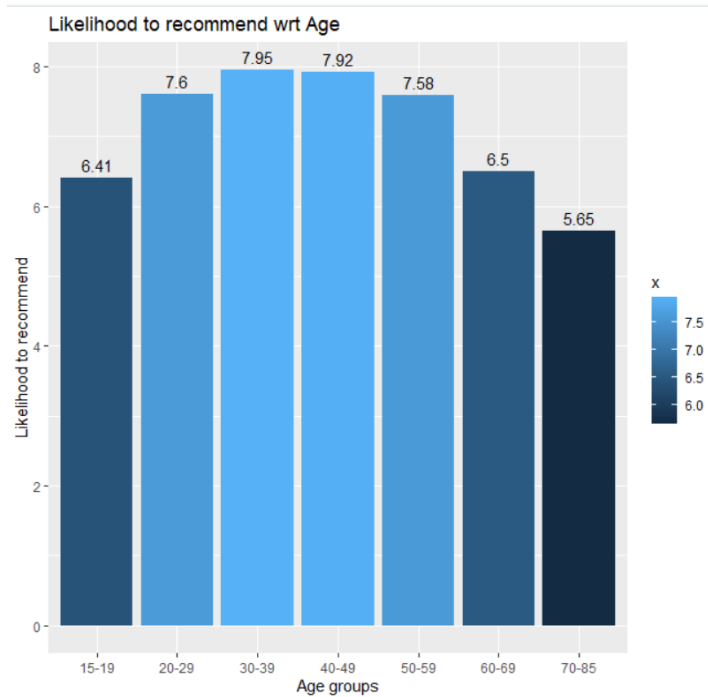
Visualization

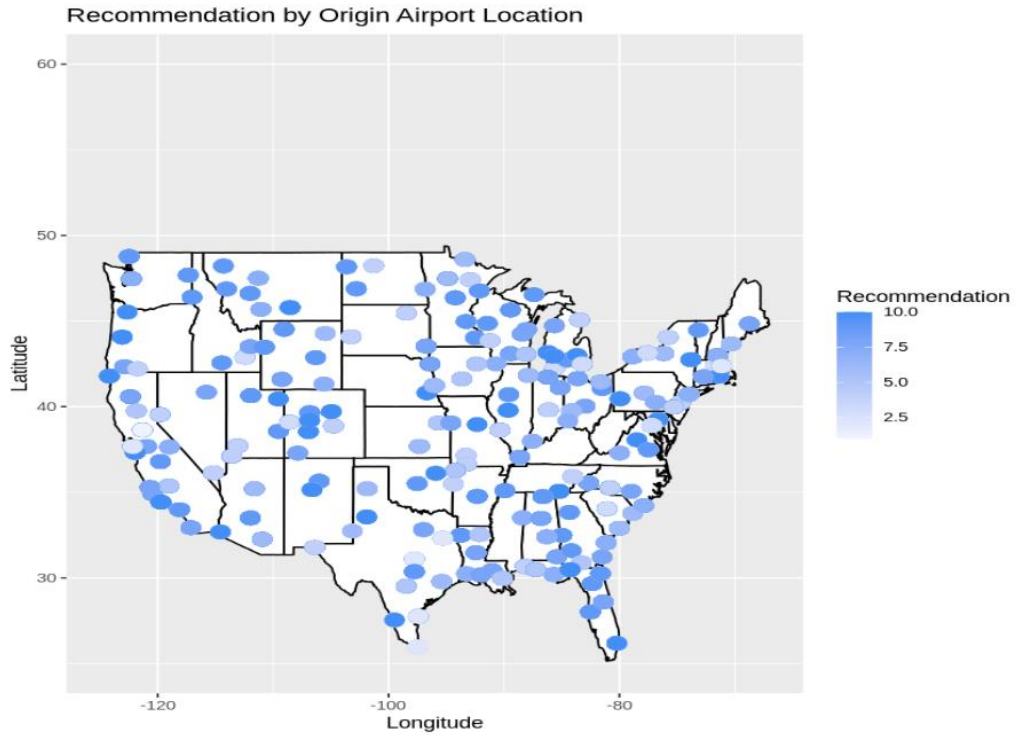


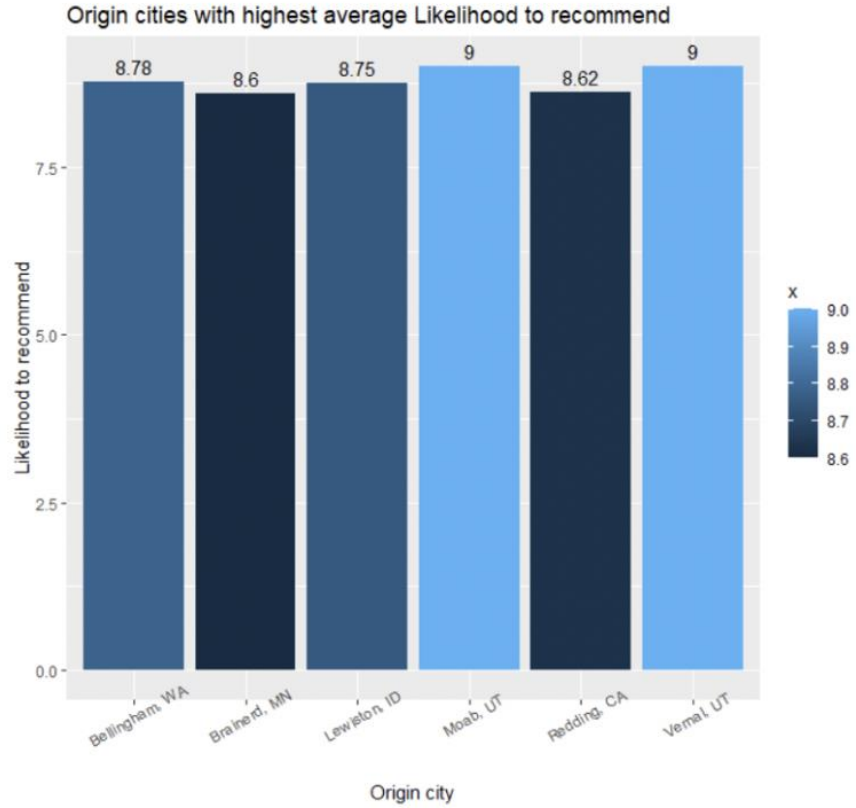

Number of Travellers wrt Gender
50000
40000
30000
20000
10000
0
Number of Travellers
Female
Male
Gender
Gender
Female
Male


Likelihood to recommend wrt gender
7.11
7.57
6
4
2
0
Likelihood to recommend
Female
Male
Gender
x
7.5
7.4
7.3
7.2


Number of travellers: Age
15000
10000
5000
0
count
15-19
20-29
30-39
40-49
50-59
60-69
70-85
Age

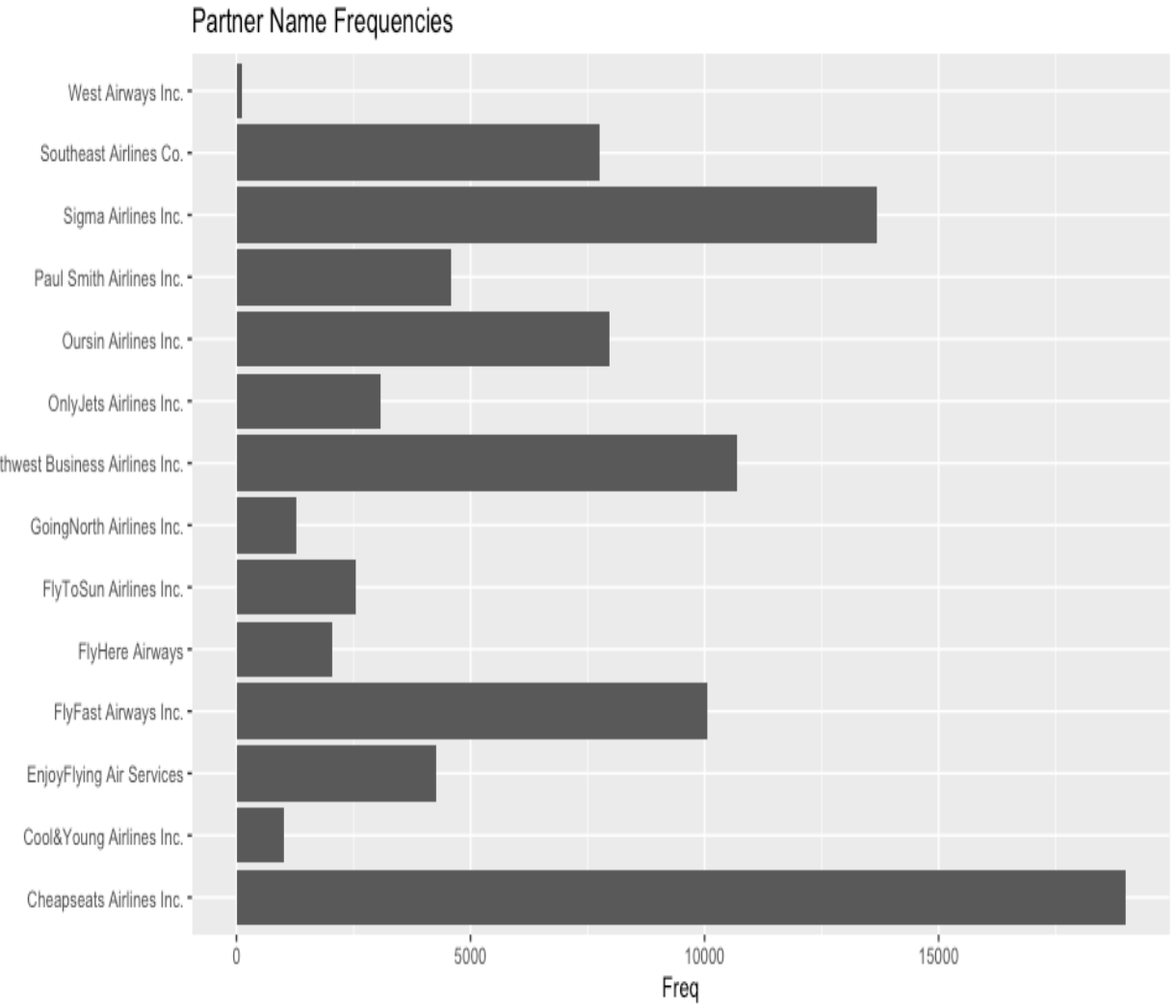

Likelihood to recommend wrt Age
6.41
7.6
7.95
7.92
7.58
6.5
5.65
8
6
4
2
0
Likelihood to recommend
15-19
20-29
30-39
40-49
50-59
60-69
70-85
Age groups
x
7.5
7.0
6.5
6.0

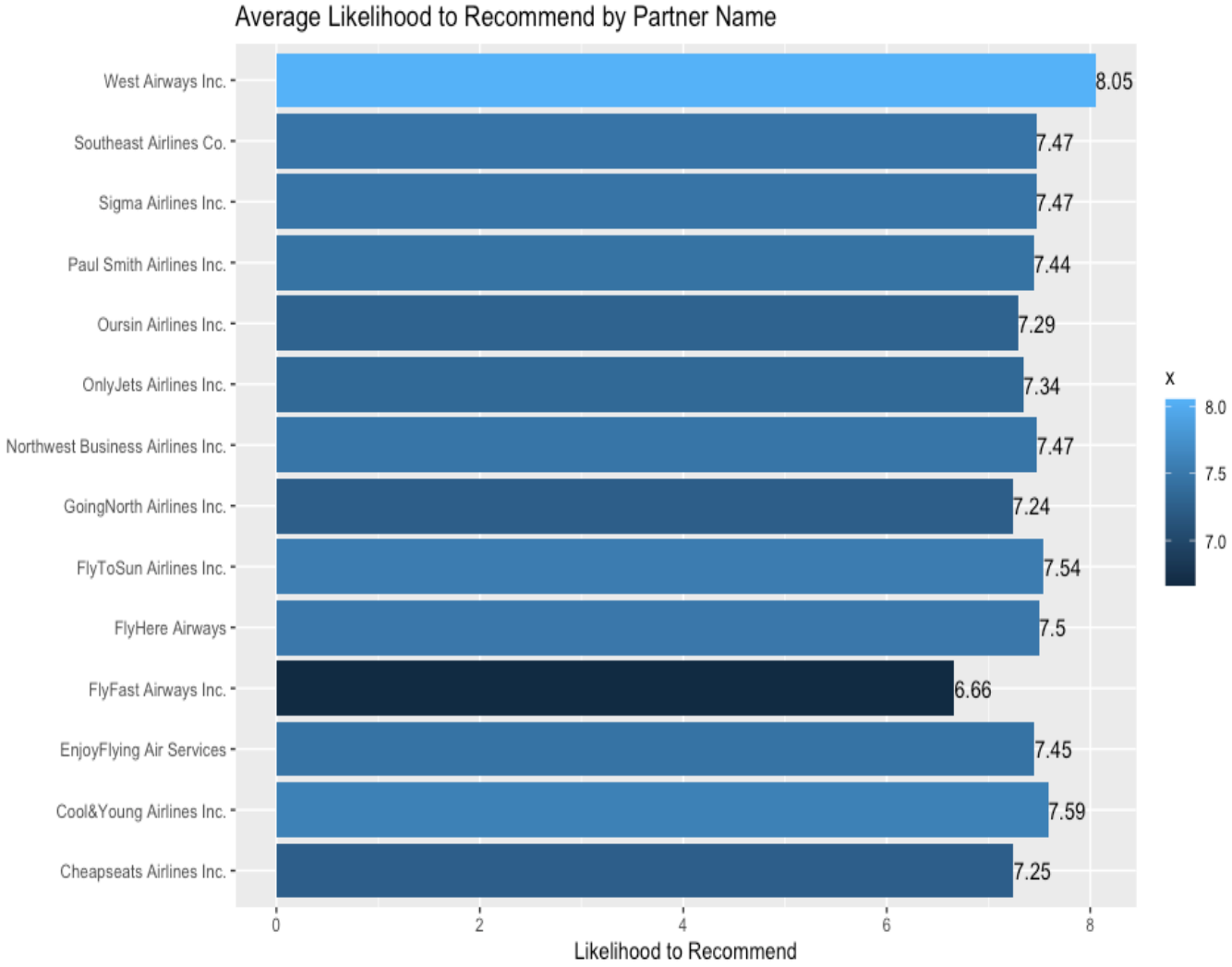

Recommendation by Origin Airport Location
60
50
40
30
Latitude
-120
-100
-80
Longitude
Recommendation
10.0
7.5
5.0
2.5

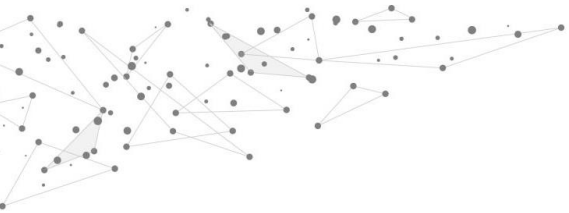

Origin cities with highest average Likelihood to recommend
8.78
8.6
8.75
9
8.62
9
7.5
5.0
2.5
0.0
Likelihood to recommend
Bellingham, WA
Brainerd, MN
Lewiston, ID
Moab, UT
Redding, CA
Vernal, UT
Origin city
x
9.0
8.9
8.8
8.7
8.6

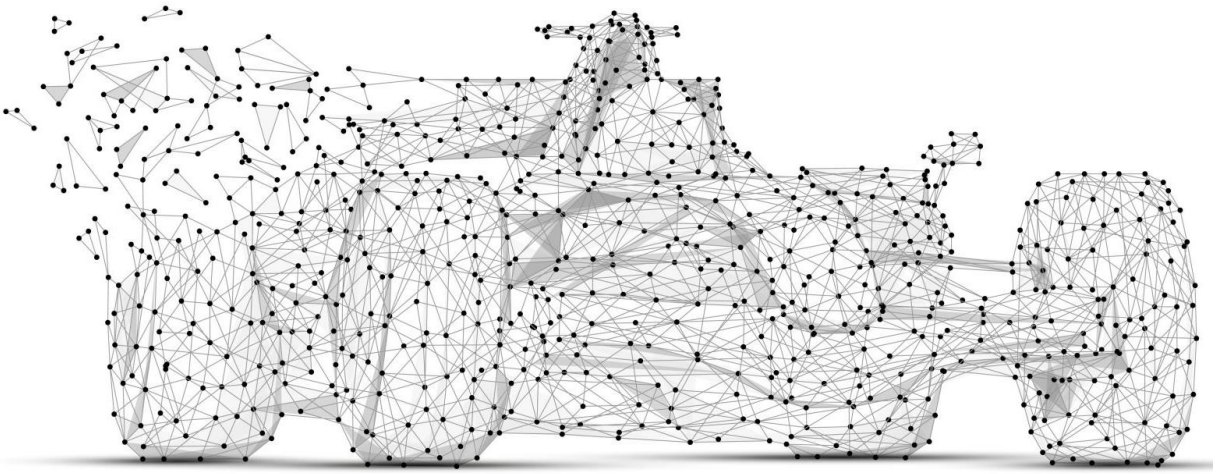
Zoom

Likelihood to recommend scores wrt Partner name







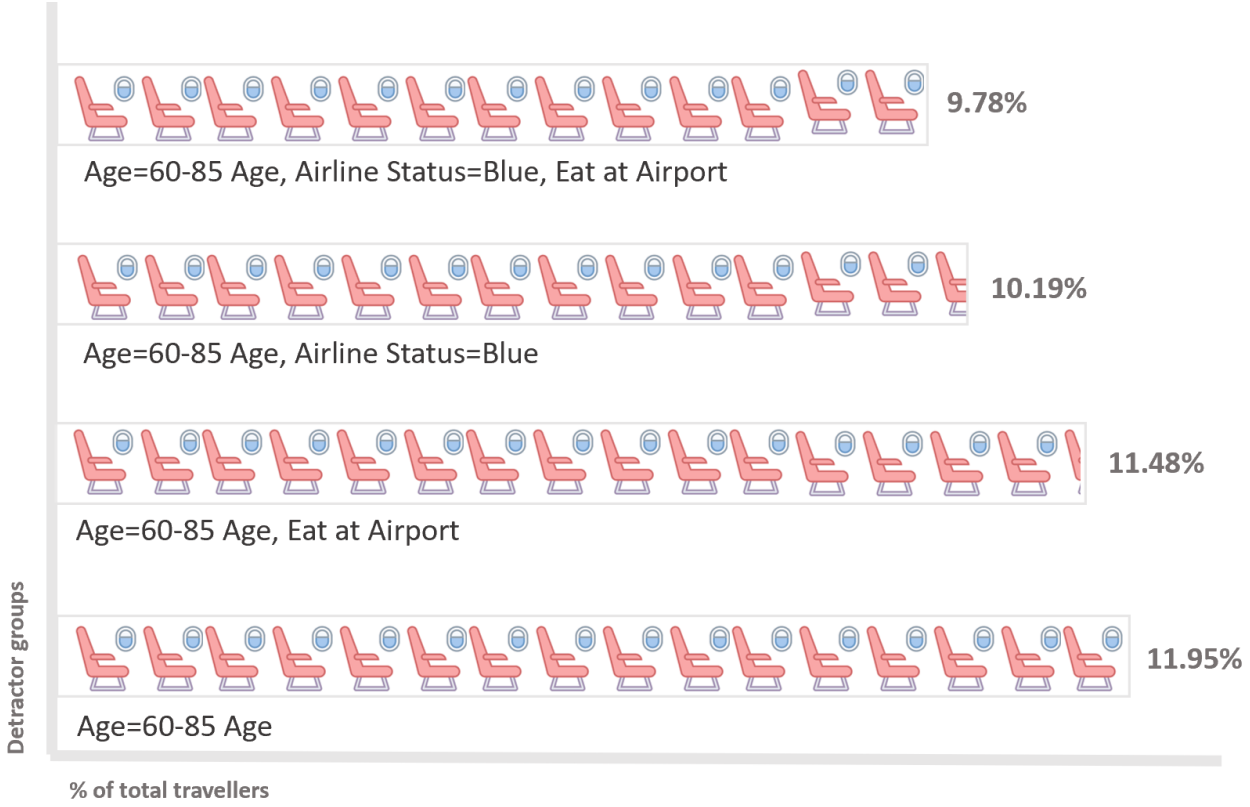


Machine Learning

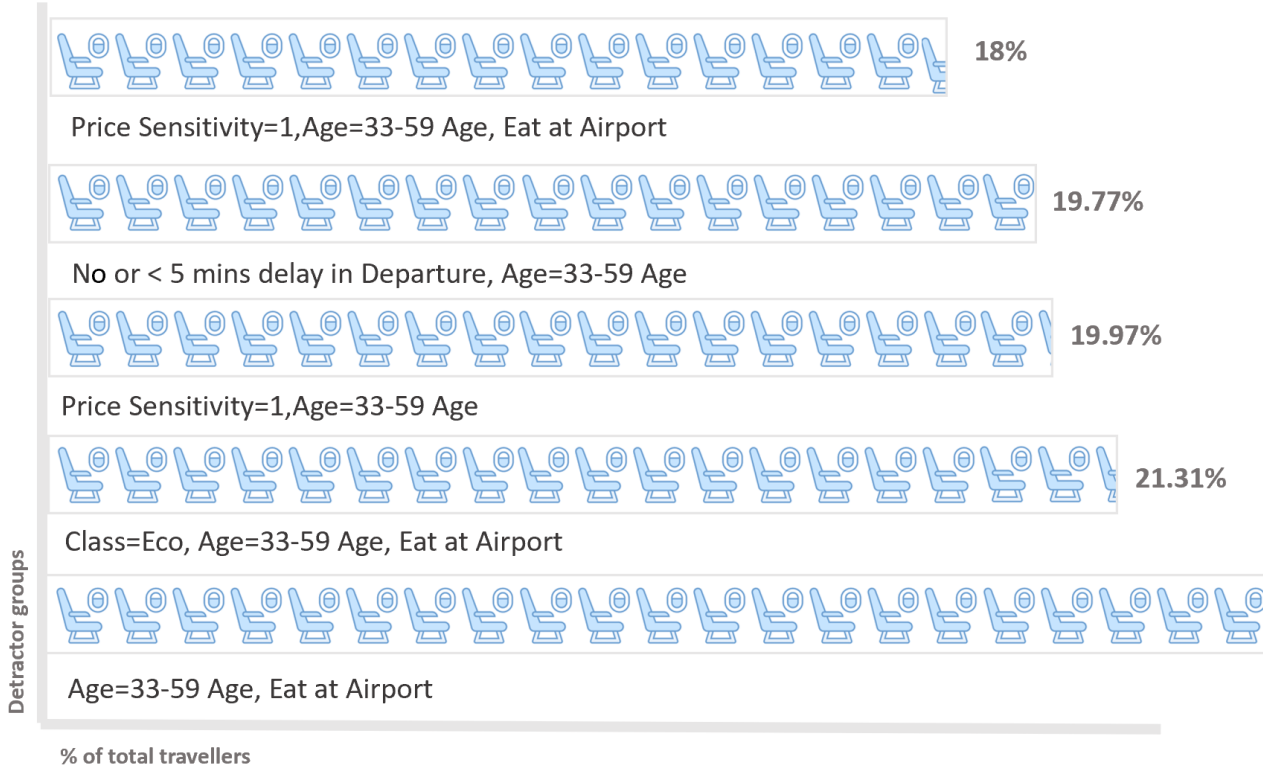
- Association Rule Mining
- Linear Modeling
- Support Vector Modeling (SVM)
- Text mining

Mining

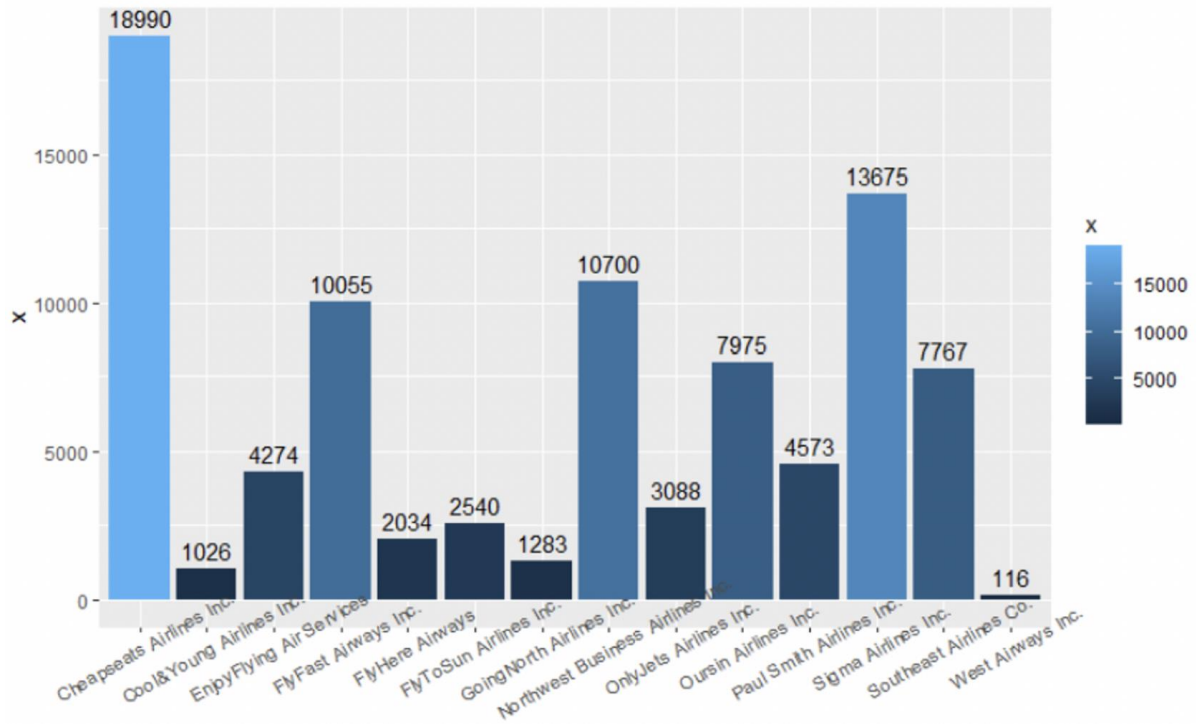
DETRACTOR GROUPS



PROMOTERS GROUPS



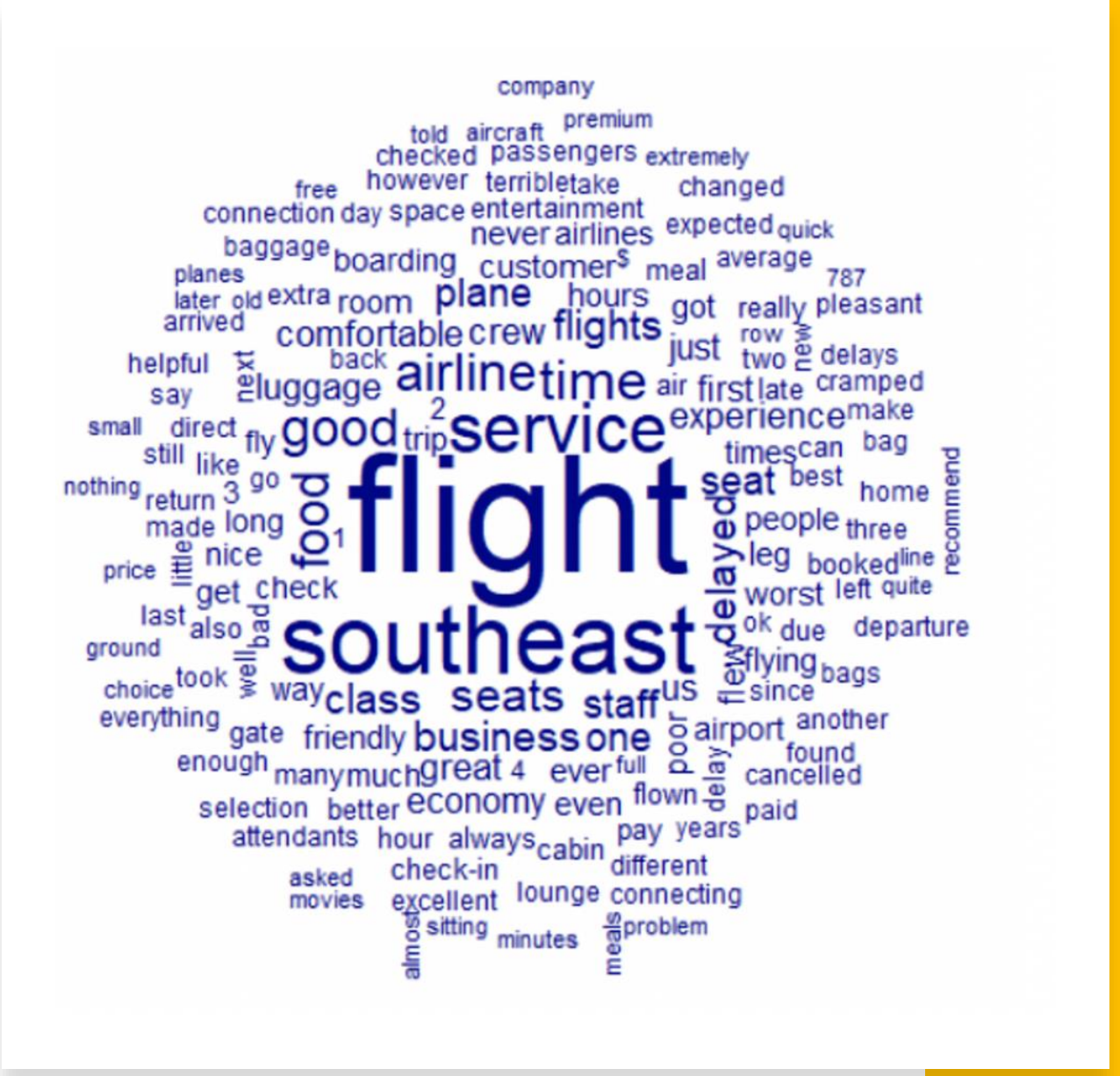
Linear Model




```
Call:
lm(formula = Likelihood.to.recommend ~ Arrival.Delay.in.Minutes +
    Departure.Delay.in.Minutes + Age + Price.Sensitivity + Flights.Per.Year +
    Loyalty + Total.Freq.Flyer.Accts + Shopping.Amount.at.Airport +
    Eating.and.Drinking.at.Airport + Flight.time.in.minutes +
    Flight.Distance, data = survey.p)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3291 -1.3076  0.4625  1.6088  6.9618

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    9.563e+00  3.717e-02 257.272  < 2e-16 ***
Arrival.Delay.in.Minutes      -8.863e-03  7.462e-04 -11.877  < 2e-16 ***
Departure.Delay.in.Minutes     3.492e-03  7.556e-04   4.622 3.81e-06 ***
Age                           -2.602e-02  4.806e-04 -54.145  < 2e-16 ***
Price.Sensitivity             -4.054e-01  1.337e-02 -30.326  < 2e-16 ***
Flights.Per.Year              -3.302e-02  7.307e-04 -45.194  < 2e-16 ***
Loyalty                       -1.406e-01  2.022e-02  -6.956 3.54e-12 ***
Total.Freq.Flyer.Accts        -6.811e-02  7.365e-03  -9.248  < 2e-16 ***
Shopping.Amount.at.Airport     3.516e-04  1.356e-04   2.592  0.00954 **
Eating.and.Drinking.at.Airport 3.089e-03  1.410e-04  21.909  < 2e-16 ***
Flight.time.in.minutes        -1.450e-03  4.879e-04  -2.972  0.00296 **
Flight.Distance                2.303e-04  5.904e-05   3.900 9.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.121 on 86246 degrees of freedom
  (1842 observations deleted due to missingness)
Multiple R-squared:  0.1083,    Adjusted R-squared:  0.1082
F-statistic: 952.6 on 11 and 86246 DF,  p-value: < 2.2e-16
```

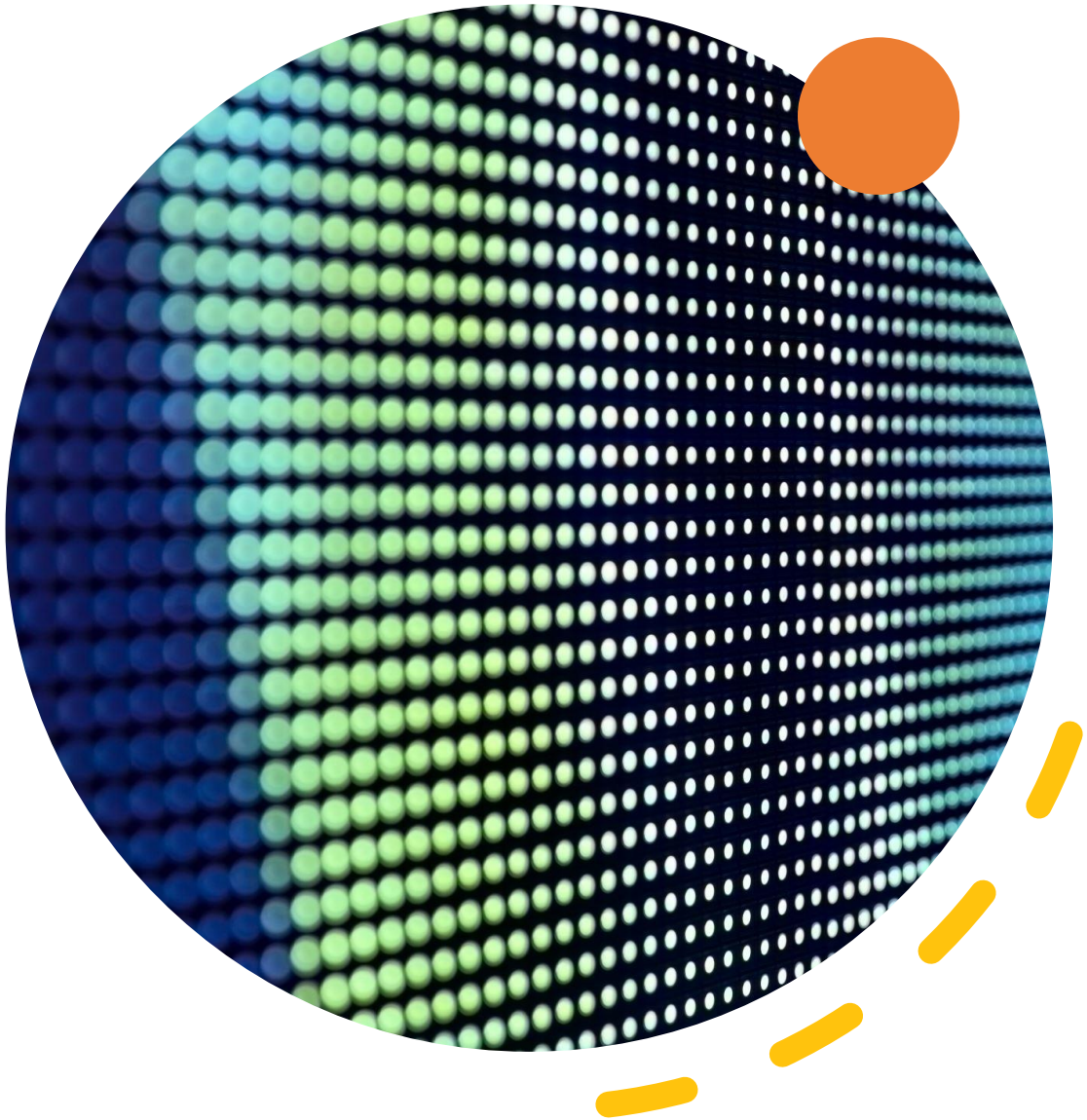
Text mining

- Positive words
- Negative words



IST 652 – Scripting for Data Analysis

- Project description
- Descriptive analysis
- Data cleaning and preparation
- Exploratory Data Analysis



Project description

- Travel is a big part of our life and during the travel booking the hotel is always the most important process. The total number of online travel bookings made each year is around 148.3 million, which generates sales of around $755 billion per year in 2020. And it continues to increase since 2014, with an average of 10% every year. But in the real world, many people are troubled by how to book a value hotel. It’s no surprise that over half the people spend more than one week researching their hotel before the holiday. Based on those questions, we make this report to solve it.

Descriptive Analysis

Count

Mean

Standard Deviation

Quartile

Max & Mean

Basic Visualization

Data Overview

- Hotel booking dataset is contains **119205 rows** with **31 columns**.

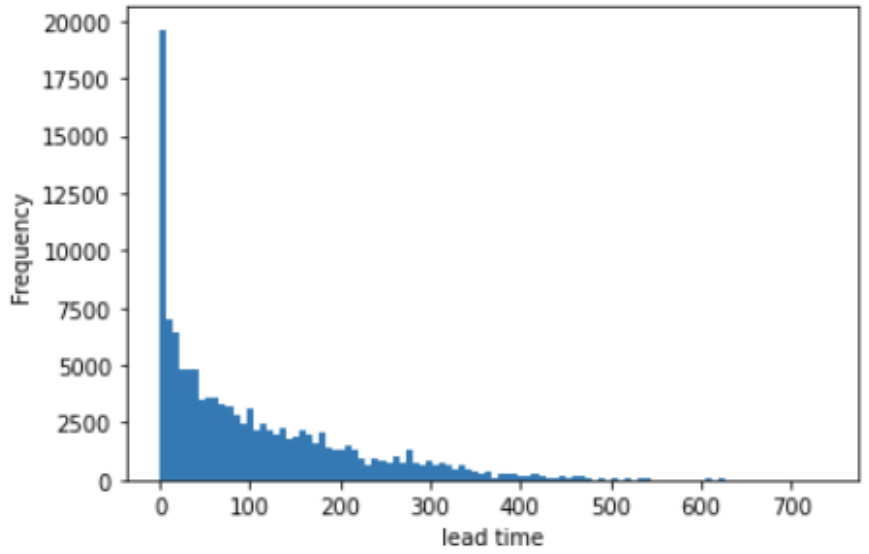
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	s
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	
...	...	...	...	...	...	...	...	...	
119200	City Hotel	0	23	2017	August	35	30	2	
119201	City Hotel	0	102	2017	August	35	31	2	
119202	City Hotel	0	34	2017	August	35	31	2	
119203	City Hotel	0	109	2017	August	35	31	2	
119204	City Hotel	0	205	2017	August	35	29	2	

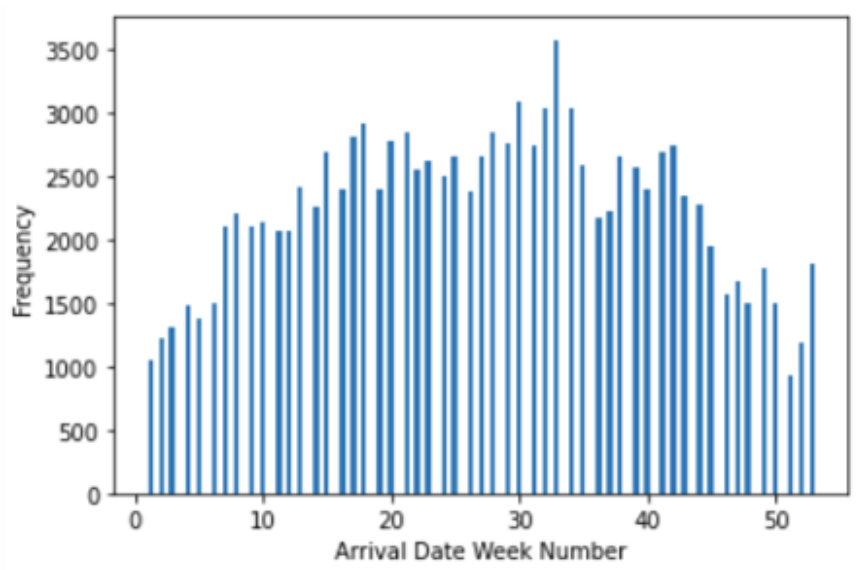
Normal varibales:

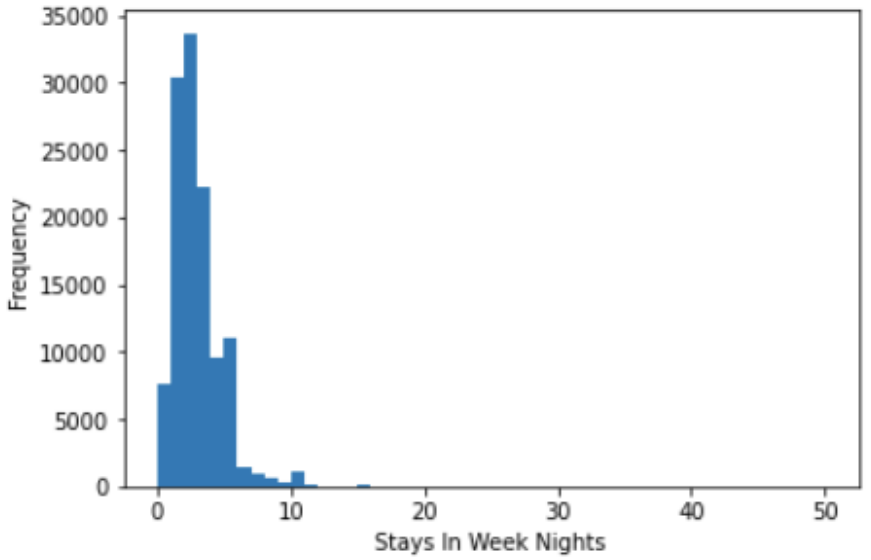
Lead Time

Arrival Date Week Number

Stays in Weekend Nights







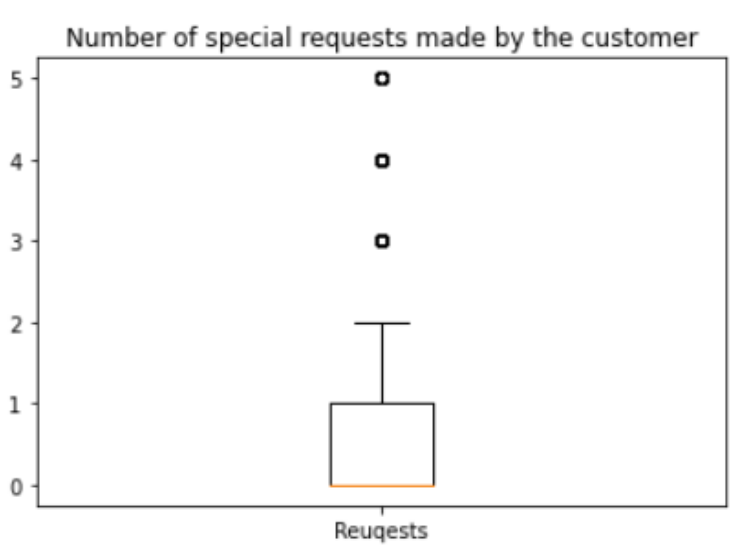
Abnormal variables:

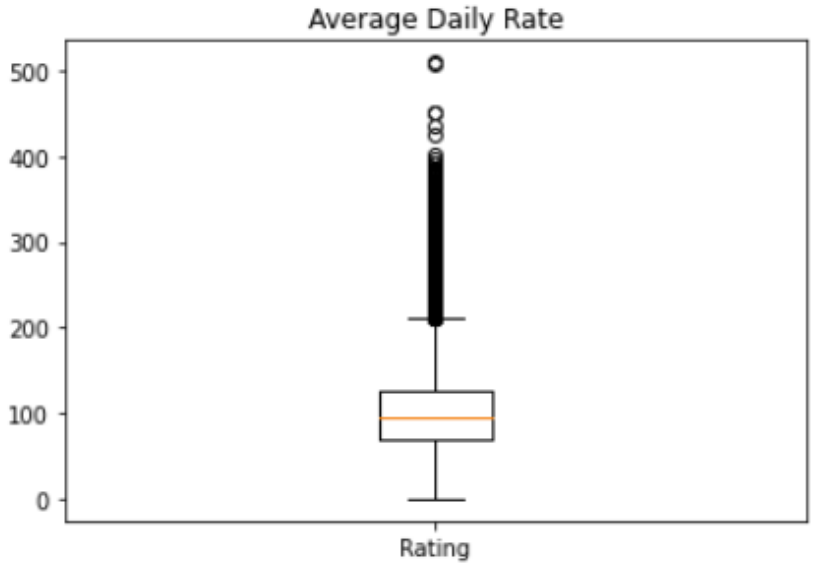
Average Daily Rate

Number of Car Parking Spaces Required by the

Number of Special Requests Made by the Custo

	adr
count	119390.000000
mean	101.831122
std	50.535790
min	-6.380000
25%	69.290000
50%	94.575000
75%	126.000000
max	5400.000000





Data Cleaning

Cleaning and Preparation

- Missing data
- Data type
- Duplicates
- Outlier
- Replacing values

Missing Value Processing

1. company: 94.31%
1. agent: 13.69%
1. children: 0.0034%
1. country: 0.41%

```
company                         0.943069
agent                           0.136862
country                         0.004087
children                        0.000034
lead_time                       0.000000
arrival_date_year               0.000000
arrival_date_month              0.000000
arrival_date_week_number        0.000000
is_canceled                     0.000000
market_segment                  0.000000
arrival_date_day_of_month       0.000000
stays_in_weekend_nights         0.000000
stays_in_week_nights            0.000000
adults                          0.000000
babies                          0.000000
meal                            0.000000
reservation_status_date         0.000000
distribution_channel            0.000000
reservation_status              0.000000
is_repeated_guest               0.000000
previous_cancellations          0.000000
previous_bookings_not_canceled  0.000000
reserved_room_type              0.000000
assigned_room_type              0.000000
booking_changes                 0.000000
deposit_type                    0.000000
days_in_waiting_list            0.000000
customer_type                   0.000000
adr                             0.000000
required_car_parking_spaces     0.000000
total_of_special_requests       0.000000
hotel                           0.000000
dtype: float64
```

Data Type

Children: float64 ------->

Children: int64

travel agency: float64 ------->
travel agency: int64

(2.5 children or agent id: 304.5)

```
Data columns (total 31 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119390 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         119390 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           119390 non-null  float64
 24  days_in_waiting_list            119390 non-null  int64
 25  customer_type                   119390 non-null  object
 26  adr                             119390 non-null  float64
 27  required_car_parking_spaces     119390 non-null  int64
 28  total_of_special_requests       119390 non-null  int64
 29  reservation_status              119390 non-null  object
 30  reservation_status_date         119390 non-null  object
```

Duplicates and Replacing

Type of meal booked. Categories are presented in standard hospitality meal packages:

Undefined/SC – no meal package; Undefined-------->SC

BB – Bed & Breakfast;

HB – Half board (breakfast and one other meal – usually dinner);

FB – Full board (breakfast, lunch and dinner)

“Since this is hotel real data, all data elements pertaining hotel or customer identification were deleted.” (Hotel booking demand datasets, 2019)

No primary key and duplicate values are allowed in this case

Outlier Handling

average daily charge of the hotel>5000

average daily charge of the hotel<0

adults+children+babies>0

adults or children or babies > 9

	count	mean	std	min	25%	50%	75%	max
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	1.0	1.0
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	2016.00	2016.000	2017.0	2017.0
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0
children	119390.0	0.103886	0.398555	0.00	0.00	0.000	0.0	10.0
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	0.0	1.0
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0
agent	119390.0	74.828319	107.141953	0.00	7.00	9.000	152.0	535.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	126.0	5400.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0

adults	2	1	2
children	0	0	10
babies	10	9	0

Visualization

Cancellation Situation & Monthly Cancellation rate

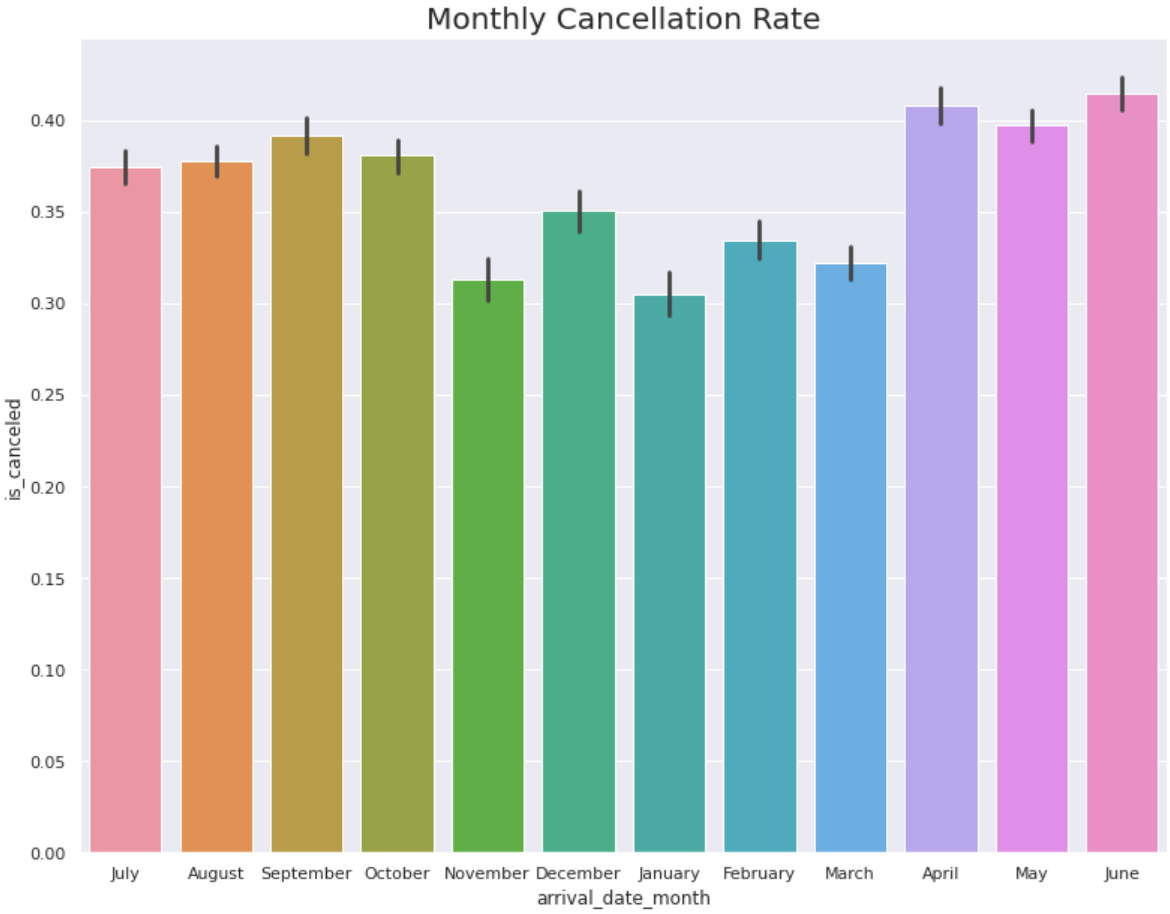
Busiest months for hotel

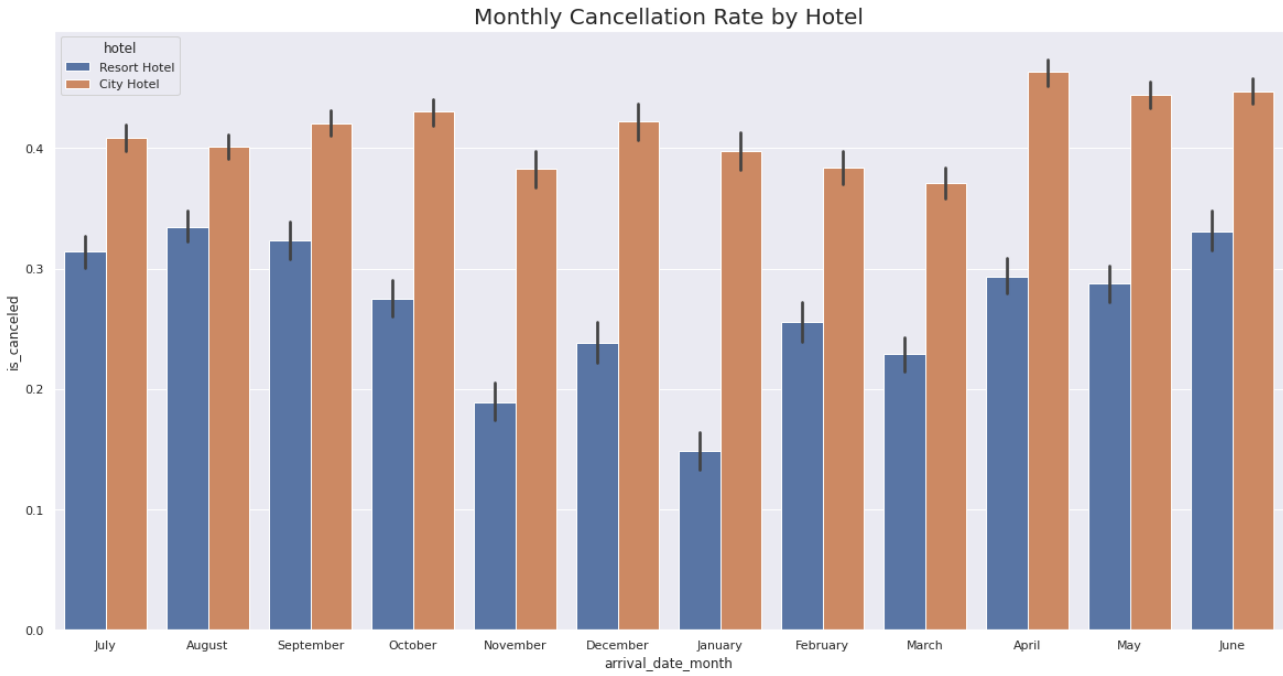
Country wise comparison

How long do people stay?

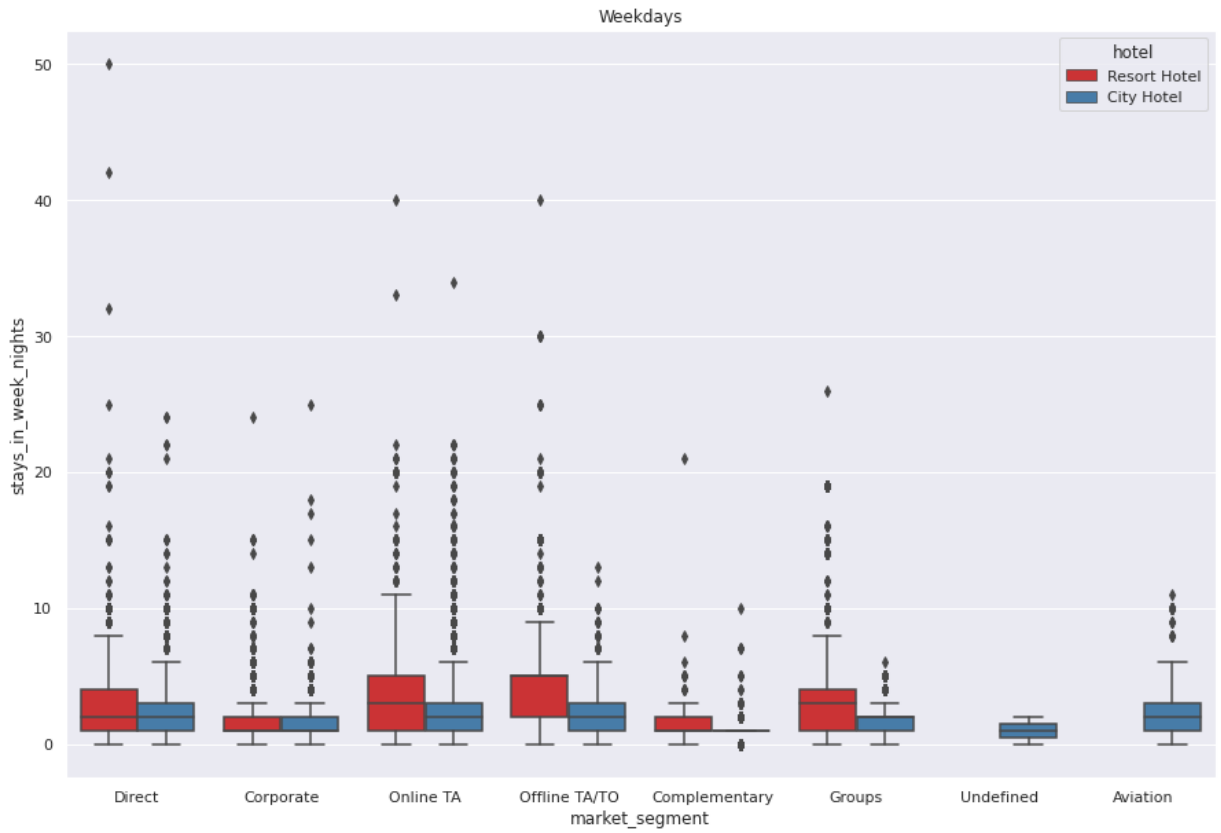
Correlation

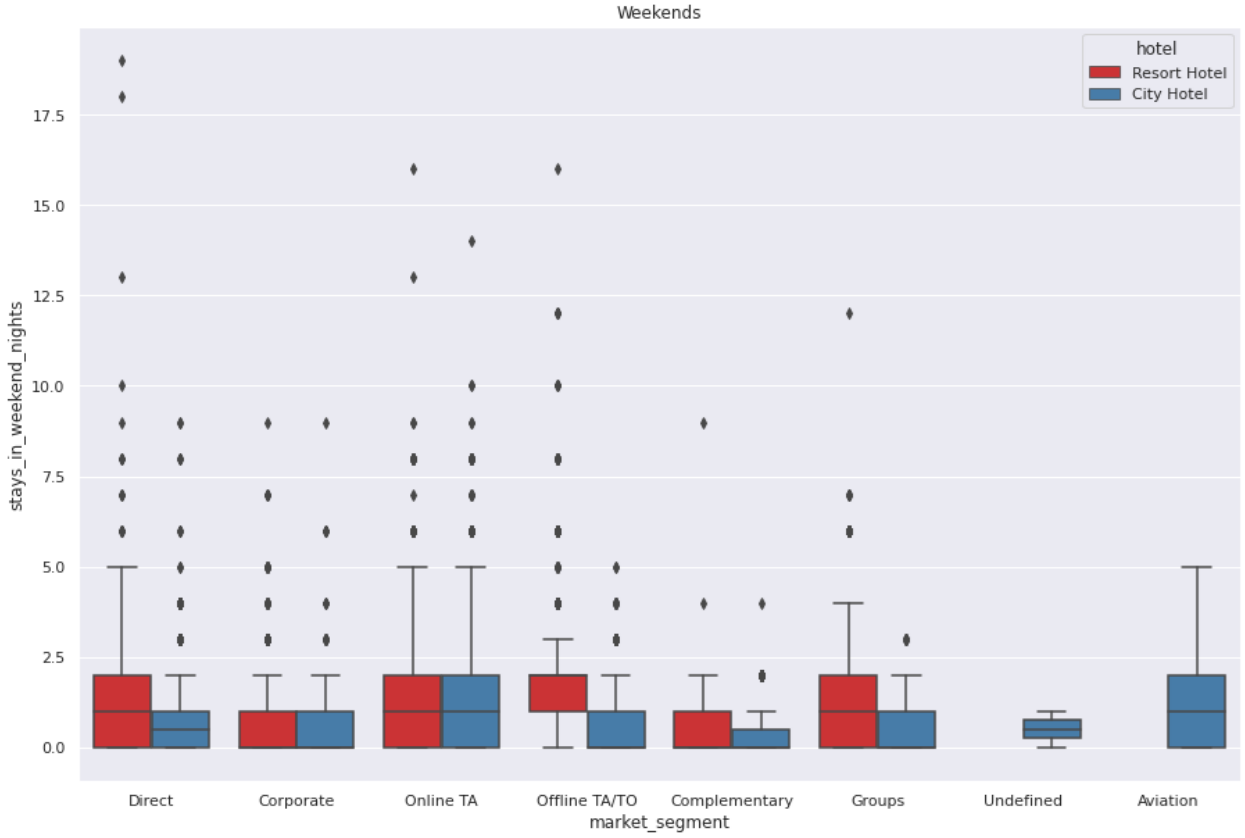
Visualization - Monthly cancellation rate



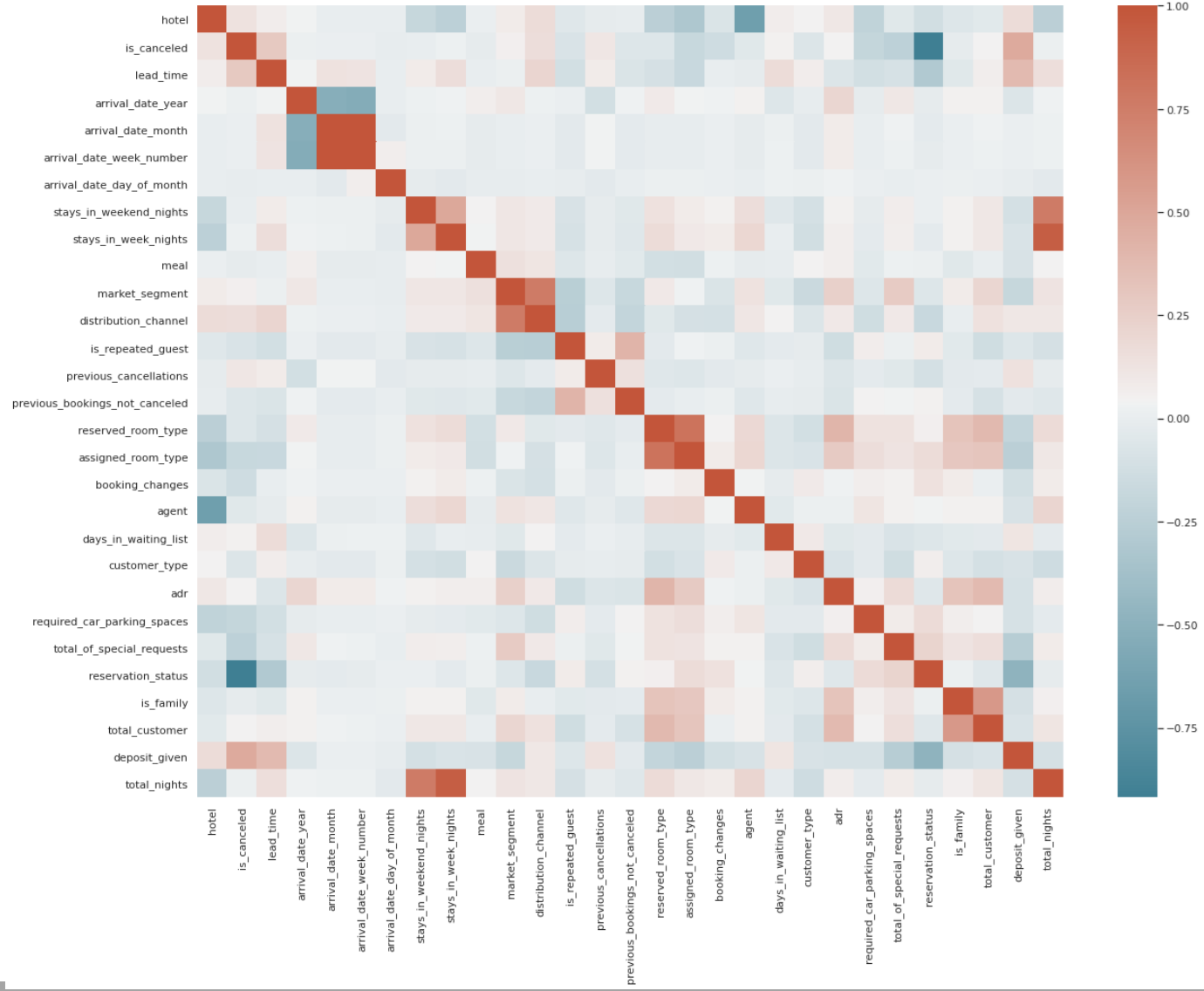


Visualization - How long do people stay?





Visualization - Correlation



is_canceled	1.000000
deposit_given	0.481501
lead_time	0.292893
distribution_channel	0.167682
hotel	0.137096
previous_cancellations	0.110142
market_segment	0.059407
days_in_waiting_list	0.054303
adr	0.047578
total_customer	0.044933
stays_in_week_nights	0.025516
total_nights	0.018533
arrival_date_year	0.016638
arrival_date_month	0.011179
arrival_date_week_number	0.008313
stays_in_weekend_nights	-0.001326
arrival_date_day_of_month	-0.005969
is_family	-0.013220
meal	-0.017226
agent	-0.046994
previous_bookings_not_canceled	-0.057358
reserved_room_type	-0.062225
customer_type	-0.068152
is_repeated_guest	-0.083721
booking_changes	-0.144857
assigned_room_type	-0.175833
required_car_parking_spaces	-0.195705
total_of_special_requests	-0.234888
reservation_status	-0.917239

Conclusion

- Resort hotels tend to have less bookings in comparison to city hotels so they need to work on their marketing strategy and promote the hotels more, especially on social media.
- Resort hotels could also reduce prices to increase booking percentages.
- May-August happens to be the busiest months but so the hotels should target more customers and try to do more business during these times.
- Although city hotels have more bookings, they also tend to have more cancellations so to prevent this they could take advance money during vacation. This would ensure most bookings to not be cancelled. They could also apply no-refund policies or make the refund policies rather strict so the customers choose not to cancel.
- It is quite clear most customers travel in pairs and bringing children or babies along are very rare so the hotels could advertise in ways that attract couples more and also business travellers.
- Most guests do not return but as these customers have already visited once, advertisements should be targeted in such ways so they are bound to return the next time they visit. The customers could also be offered special benefits if they do return to stay.

IST 707 – Data Analytics

- Project overview
- Data preprocessing
- Data mining and modeling
- Results

Overview

Goal:

Three different machine learning algorithms KNN, SVM and Apriori will be used in this data mining project.

PREDICTION

ASSOCIATION RULES

Tool: R Studio, Weka

Dataset:

This dataset contains a list of video games with sales greater than 100,000 copies.

Size: 2 MB

Fields:

Rank, Name , Platform , Year , Genre , Publisher , NA_Sales , EU_Sales , JP_Sales , Other_Sales , Global_Sales

Dataset:

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sale	Global_Sal
1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	Super Mar	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	Mario Kart	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	Wii Sports	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
5	Pokemon I	GB	1996	Role-Playi	Nintendo	11.27	8.89	10.22	1	31.37
6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
7	New Super	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
9	New Super	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31

Preparation

1. Missing values
2. Data type
3. Outlier
4. Rank
5. Training and test sets (60%-40%)
6. Descriptive analysis

```
# Data type
str(vgsales)
vgsales$Year<-as.integer(vgsales$Year)
table(is.na(vgsales$Year))
unique(vgsales$Year)
vgsales<- na.omit(vgsales)
table(is.na(vgsales$Year))
table(is.na(vgsales))
```

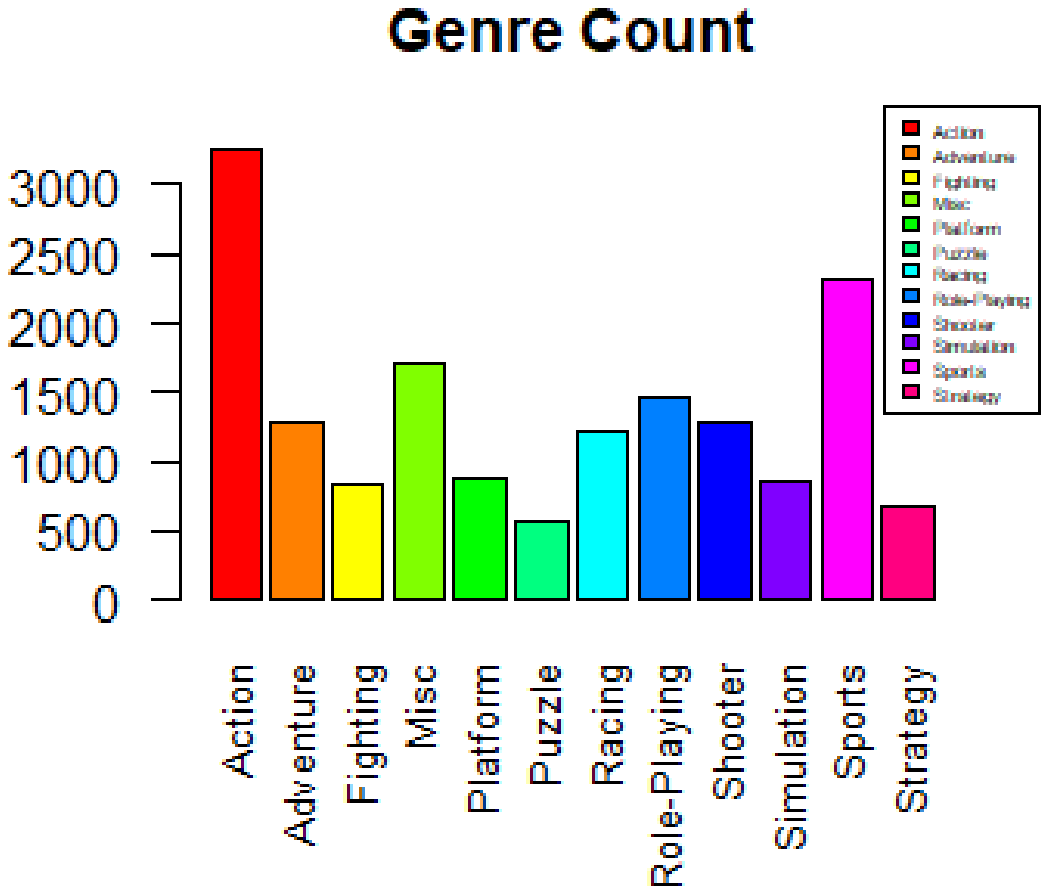
```
> summary(sales)
   Group.1                x
 Length:577         Min.   :   0.01
 Class :character   1st Qu.:   0.07
 Mode  :character   Median :   0.32
                    Mean   :  15.29
                    3rd Qu.:   1.64
                    Max.   :1784.43
```

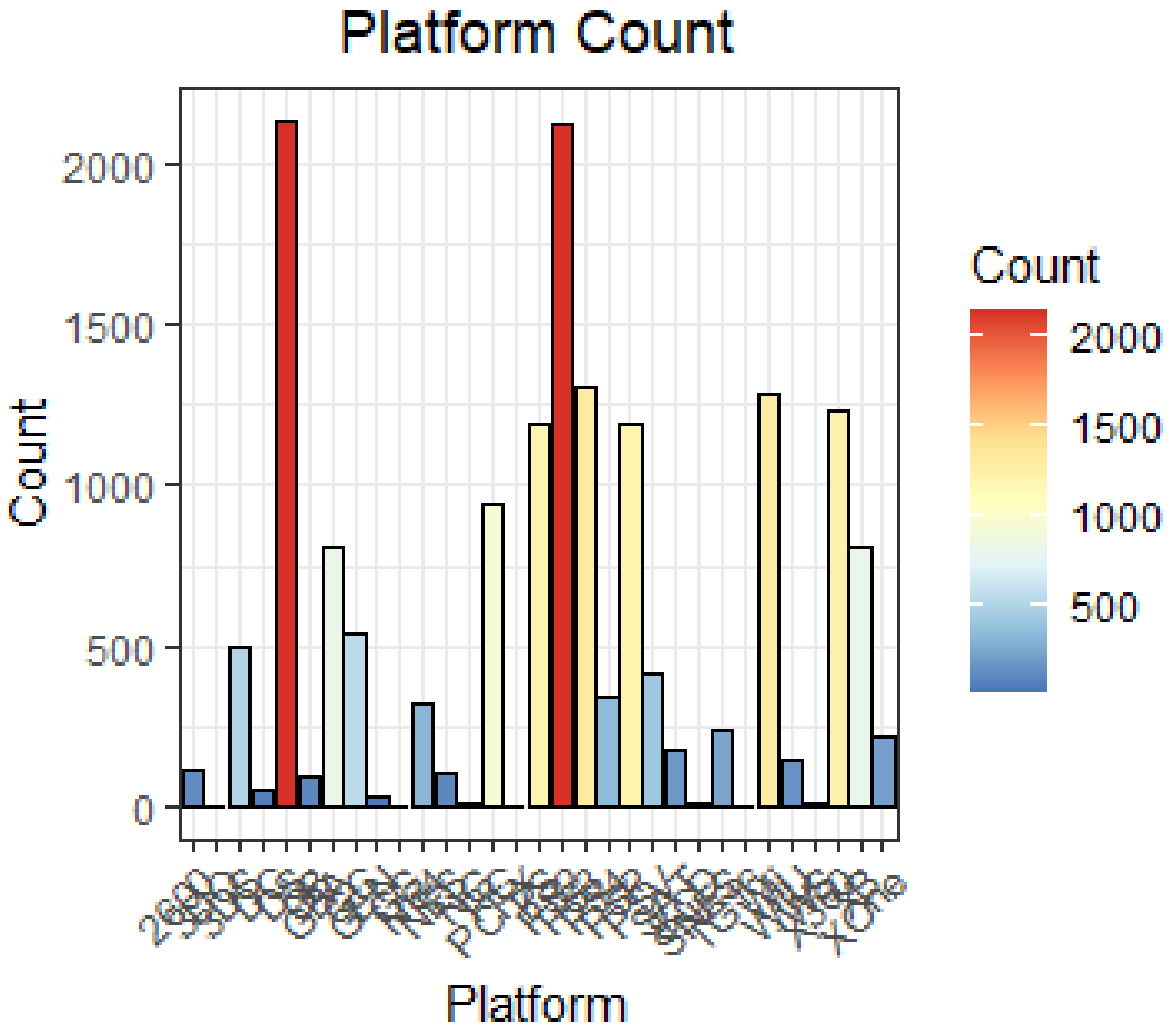
```
# Outliers
unique(vgsales$Platform)
length(unique(vgsales$Platform))

unique(vgsales$Genre)
length(unique(vgsales$Genre))

unique(vgsales$Publisher)
length(unique(vgsales$Publisher))

unique(vgsales$Year)
length(unique(vgsales$Year))
min(vgsales$Year)
max(vgsales$Year)
boxplot(vgsales$Year)

vgsales[which(vgsales$Year==1980),]
vgsales[which(vgsales$Year==2017),]
vgsales[which(vgsales$Year==2020),] # Outlier

vgsales<- vgsales[-which(vgsales$Year==2020),]
```

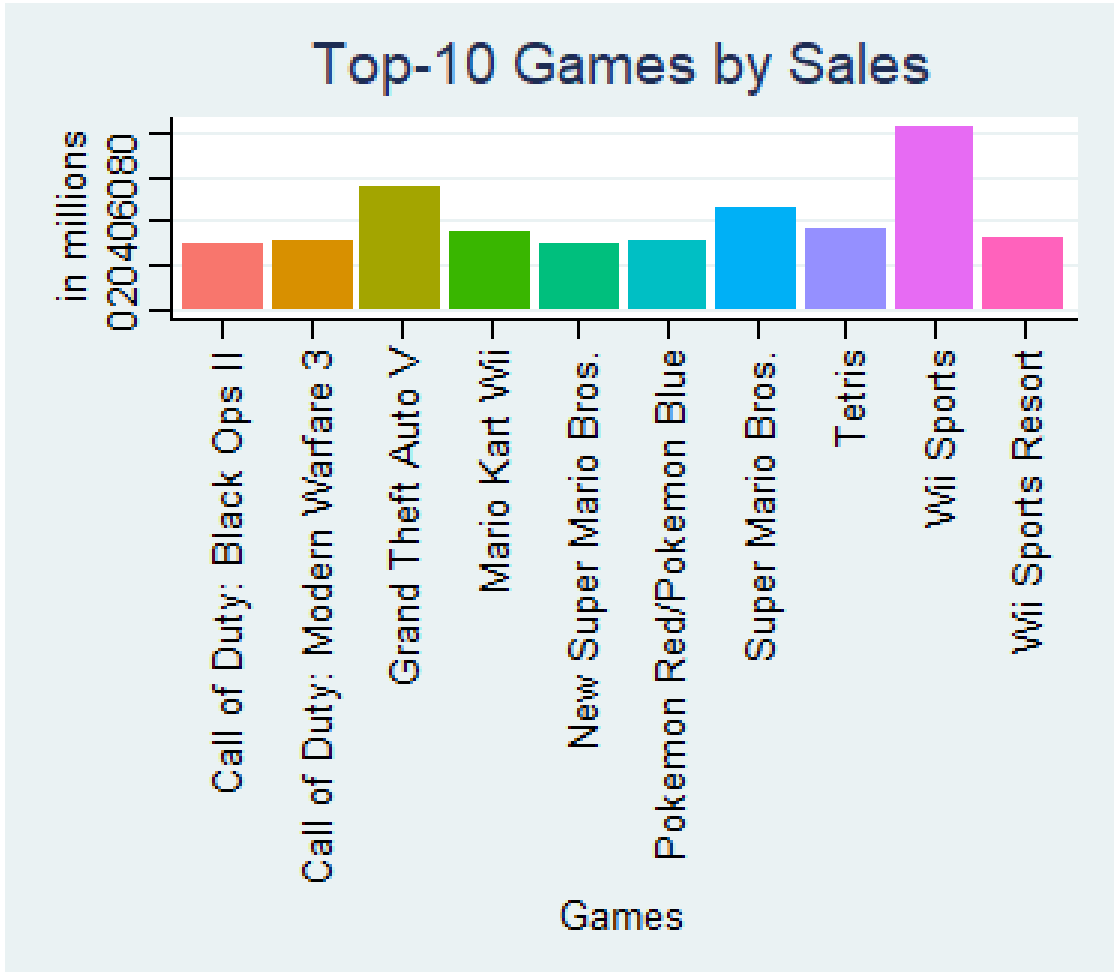
```
> summary(vgsales)
      Rank            Name            Platform            Year         Genre          Publisher          NA_Sales
 Min.   :    1   Length:16324       Length:16324       Min.   :1980   Length:16324       Length:16324       Min.   : 0.0000
 1st Qu.: 4082   Class :character   Class :character   1st Qu.:2003   Class :character   Class :character   1st Qu.: 0.0000
 Median : 8162   Mode  :character   Mode  :character   Median :2007   Mode  :character   Mode  :character   Median : 0.0800
 Mean   : 8162                                         Mean   :2006                                         Mean   : 0.2655
 3rd Qu.:12243                                         3rd Qu.:2010                                         3rd Qu.: 0.2400
 Max.   :16324                                         Max.   :2017                                         Max.   :41.4900
    EU_Sales          JP_Sales         Other_Sales        Global_Sales
 Min.   : 0.0000   Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.0100
 1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.: 0.0600
 Median : 0.0200   Median : 0.00000   Median : 0.01000   Median : 0.1700
 Mean   : 0.1476   Mean   : 0.07867   Mean   : 0.04833   Mean   : 0.5403
 3rd Qu.: 0.1100   3rd Qu.: 0.04000   3rd Qu.: 0.04000   3rd Qu.: 0.4800
 Max.   :29.0200   Max.   :10.22000   Max.   :10.57000   Max.   :82.7400
```

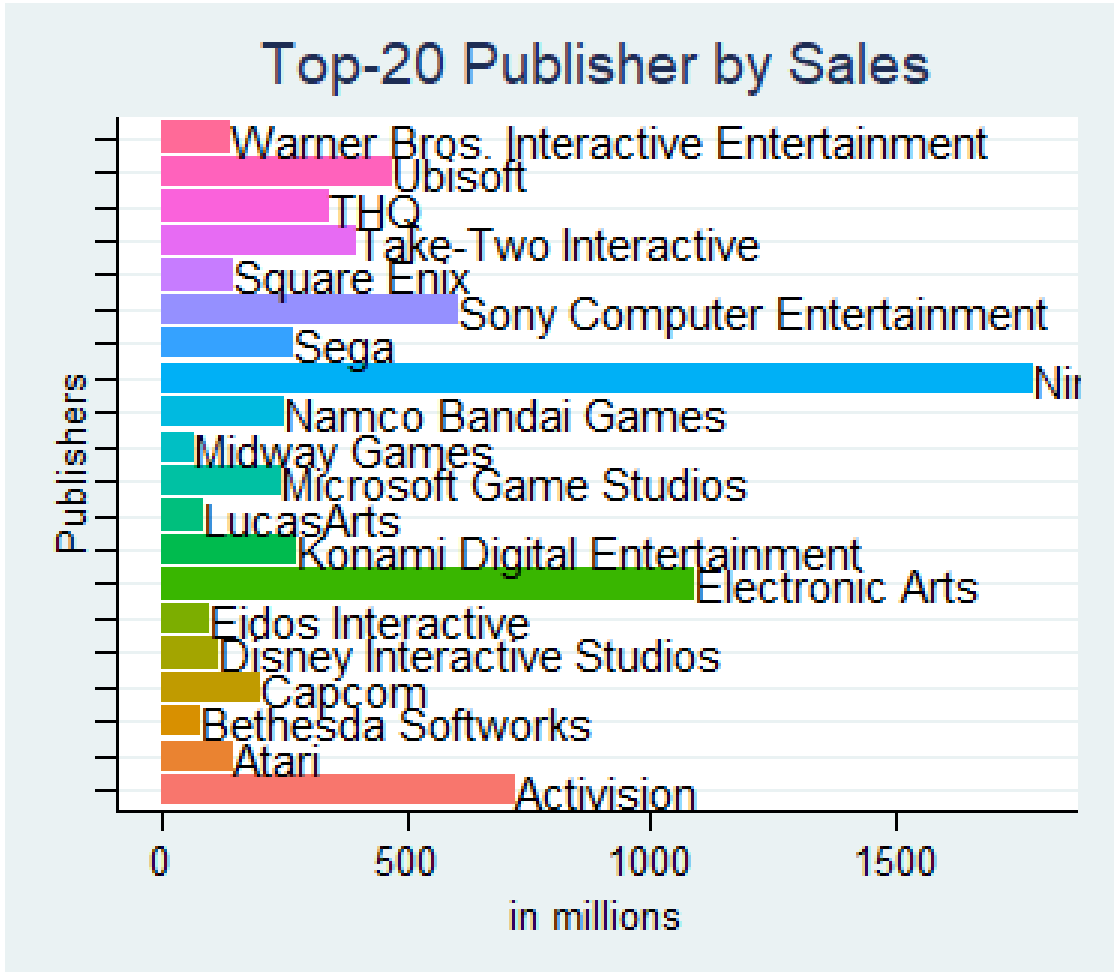
```
> str(vgsales)
'data.frame':   16324 obs. of  11 variables:
 $ Rank        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name        : chr  "Wii Sports" "Super Mario Bros." "Mario Kart Wii" "Wii Sports Resort" ...
 $ Platform    : chr  "Wii" "NES" "Wii" "Wii" ...
 $ Year        : int  2006 1985 2008 2009 1996 1989 2006 2006 2009 1984 ...
 $ Genre       : chr  "Sports" "Platform" "Racing" "Sports" ...
 $ Publisher   : chr  "Nintendo" "Nintendo" "Nintendo" "Nintendo" ...
 $ NA_Sales    : num  41.5 29.1 15.8 15.8 11.3 ...
 $ EU_Sales    : num  29.02 3.58 12.88 11.01 8.89 ...
 $ JP_Sales    : num  3.77 6.81 3.79 3.28 10.22 ...
 $ Other_Sales : num  8.46 0.77 3.31 2.96 1 0.58 2.9 2.85 2.26 0.47 ...
 $ Global_Sales: num  82.7 40.2 35.8 33 31.4 ...
 - attr(*, "na.action")= 'omit' Named int [1:273] 180 378 432 471 608 625 650 653 712 783 ...
  ..- attr(*, "names")= chr [1:273] "180" "378" "432" "471" ...
```

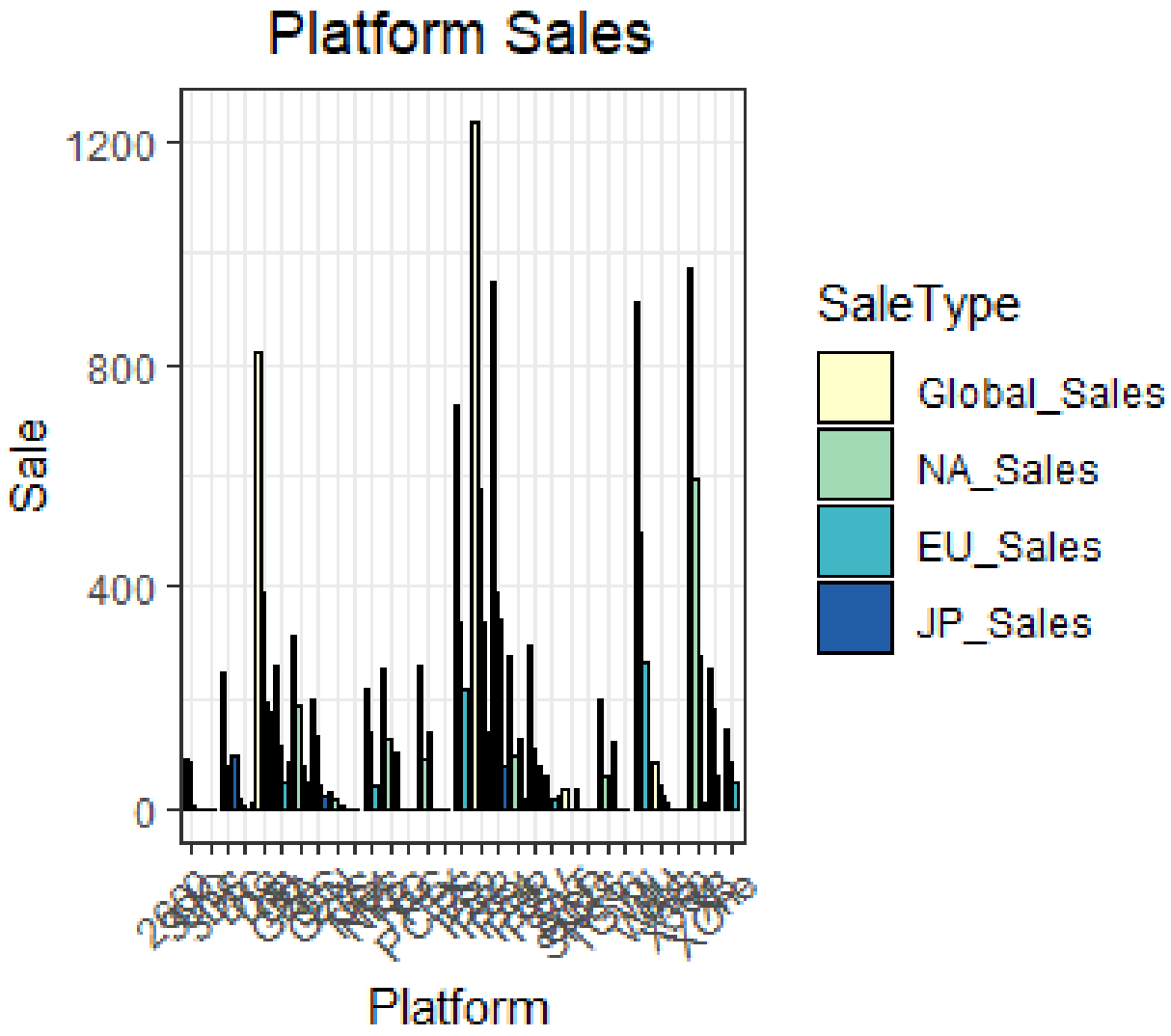
Visualization






Top-10 Games by Sales
in millions
0
20
40
60
80
Call of Duty: Black Ops II
Call of Duty: Modern Warfare 3
Grand Theft Auto V
Mario Kart Wii
New Super Mario Bros.
Pokemon Red/Pokemon Blue
Super Mario Bros.
Tetris
Wii Sports
Wii Sports Resort
Games


Top-20 Publisher by Sales
Publishers
Warner Bros. Interactive Entertainment
Ubisoft
THQ
Take-Two Interactive
Square Enix
Sony Computer Entertainment
Sega
Nin
Namco Bandai Games
Midway Games
Microsoft Game Studios
LucasArts
Konami Digital Entertainment
Electronic Arts
Eidos Interactive
Disney Interactive Studios
Capcom
Bethesda Softworks
Atari
Activision
0
500
1000
1500
in millions

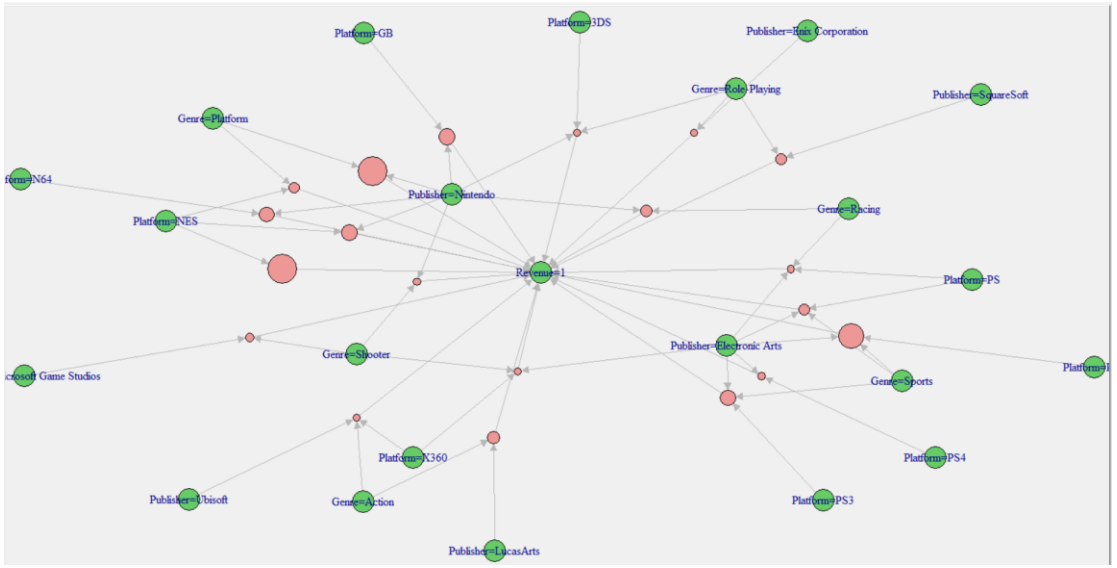


1. Nintendo
2. X360
3. ps3
4. Wii
5. Ds

Apriori

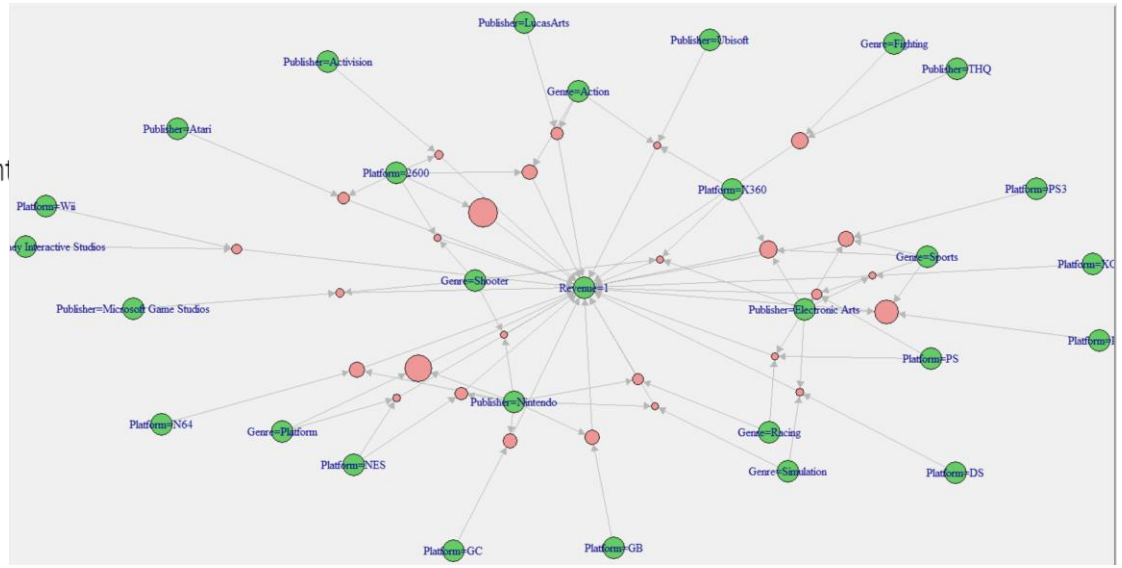
Global Sales

```
> inspect(goodrules[1:10])
     lhs                                                  rhs          support     confidence coverage    lift     count
[1]  {Platform=NES,Publisher=Nintendo}                 => {Revenue=1} 0.002756677 1.0000000  0.002756677 4.052632 45
[2]  {Platform=NES,Genre=Platform}                     => {Revenue=1} 0.001715266 1.0000000  0.001715266 4.052632 28
[3]  {Platform=NES}                                    => {Revenue=1} 0.005452095 0.9081633  0.006003431 3.680451 89
[4]  {Genre=Racing,Publisher=Nintendo}                 => {Revenue=1} 0.001899044 0.8378378  0.002266601 3.395448 31
[5]  {Platform=3DS,Genre=Role-Playing,Publisher=Nintendo} => {Revenue=1} 0.001041411 0.8095238  0.001286449 3.280702 17
[6]  {Genre=Platform,Publisher=Nintendo}               => {Revenue=1} 0.005390836 0.7927928  0.006799804 3.212897 88
[7]  {Genre=Shooter,Publisher=Microsoft Game Studios}  => {Revenue=1} 0.001347709 0.7857143  0.001715266 3.184211 22
[8]  {Platform=GB,Publisher=Nintendo}                  => {Revenue=1} 0.002879196 0.7833333  0.003675570 3.174561 47
[9]  {Platform=X360,Genre=Shooter,Publisher=Electronic Arts} => {Revenue=1} 0.001041411 0.7727273  0.001347709 3.131579 17
[10] {Platform=PS4,Publisher=Electronic Arts}          => {Revenue=1} 0.001225190 0.7692308  0.001592747 3.117409 20
> plot(goodrules,method="graph",interactive=TRUE,shading=NA)
```



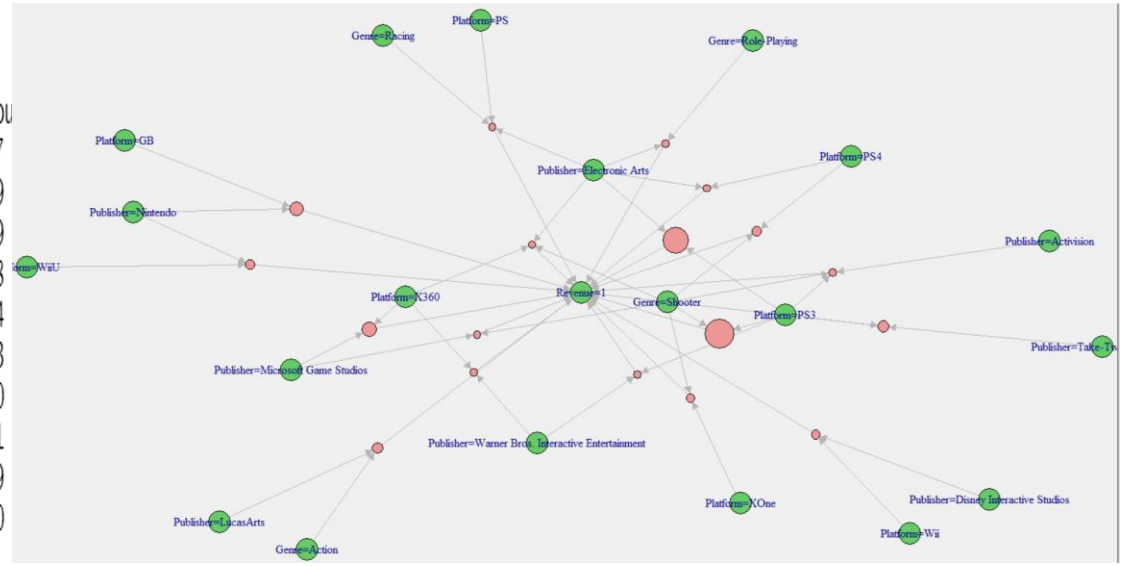
NA

```
     lhs                                                  rhs          support     confidence coverage    lift     count
[1]  {Platform=2600,Publisher=Activision}              => {Revenue=1} 0.001408968 0.9200000  0.001531487 3.751706 23
[2]  {Platform=2600,Genre=Shooter}                     => {Revenue=1} 0.001163930 0.8636364  0.001347709 3.521859 19
[3]  {Genre=Shooter,Publisher=Microsoft Game Studios}  => {Revenue=1} 0.001408968 0.8214286  0.001715266 3.349738 23
[4]  {Platform=X360,Genre=Shooter,Publisher=Electronic Arts} => {Revenue=1} 0.001102671 0.8181818  0.001347709 3.336498 18
[5]  {Genre=Racing,Publisher=Nintendo}                 => {Revenue=1} 0.001837785 0.8108108  0.002266601 3.306439 30
[6]  {Platform=2600}                                   => {Revenue=1} 0.005758393 0.8103448  0.007106101 3.304539 94
[7]  {Platform=NES,Publisher=Nintendo}                 => {Revenue=1} 0.002205342 0.8000000  0.002756677 3.262353 36
[8]  {Platform=2600,Genre=Action}                      => {Revenue=1} 0.002695418 0.8000000  0.003369272 3.262353 44
[9]  {Platform=N64,Publisher=Nintendo}                 => {Revenue=1} 0.002879196 0.7966102  0.003614310 3.248530 47
[10] {Genre=Platform,Publisher=Nintendo}               => {Revenue=1} 0.005268317 0.7747748  0.006799804 3.159486 86
```



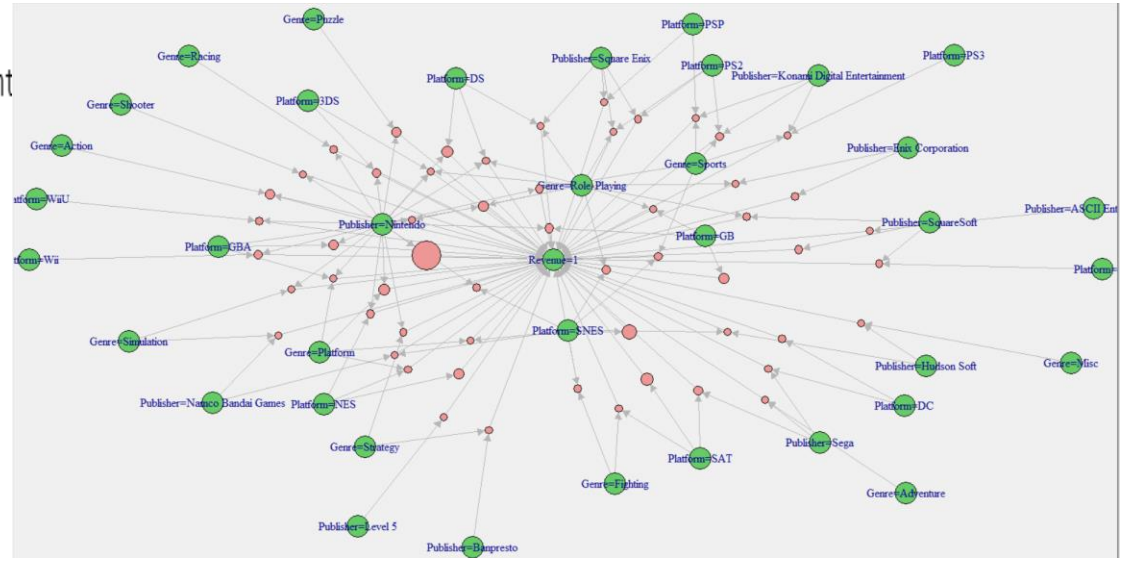
EU

```
> inspect(goodrules[1:10])
     lhs                                                                rhs          support     confidence coverage    lift     cou
[1]  {Platform=X360,Genre=Shooter,Publisher=Electronic Arts}          => {Revenue=1} 0.001041411 0.7727273  0.001347709 3.658353 17
[2]  {Platform=PS4,Publisher=Electronic Arts}                         => {Revenue=1} 0.001163930 0.7307692  0.001592747 3.459709 19
[3]  {Platform=PS3,Genre=Shooter,Publisher=Activision}                => {Revenue=1} 0.001163930 0.7307692  0.001592747 3.459709 19
[4]  {Platform=PS,Genre=Racing,Publisher=Electronic Arts}             => {Revenue=1} 0.001102671 0.7200000  0.001531487 3.408724 18
[5]  {Platform=PS4,Genre=Shooter}                                     => {Revenue=1} 0.001470228 0.7058824  0.002082823 3.341886 24
[6]  {Platform=XOne,Genre=Shooter}                                    => {Revenue=1} 0.001408968 0.6969697  0.002021563 3.299691 23
[7]  {Platform=PS3,Publisher=Warner Bros. Interactive Entertainment}  => {Revenue=1} 0.001225190 0.6896552  0.001776525 3.265061 20
[8]  {Platform=GB,Publisher=Nintendo}                                 => {Revenue=1} 0.002511639 0.6833333  0.003675570 3.235131 41
[9]  {Genre=Shooter,Publisher=Microsoft Game Studios}                 => {Revenue=1} 0.001163930 0.6785714  0.001715266 3.212587 19
[10] {Platform=X360,Publisher=Warner Bros. Interactive Entertainment} => {Revenue=1} 0.001225190 0.6451613  0.001899044 3.054412 20
```

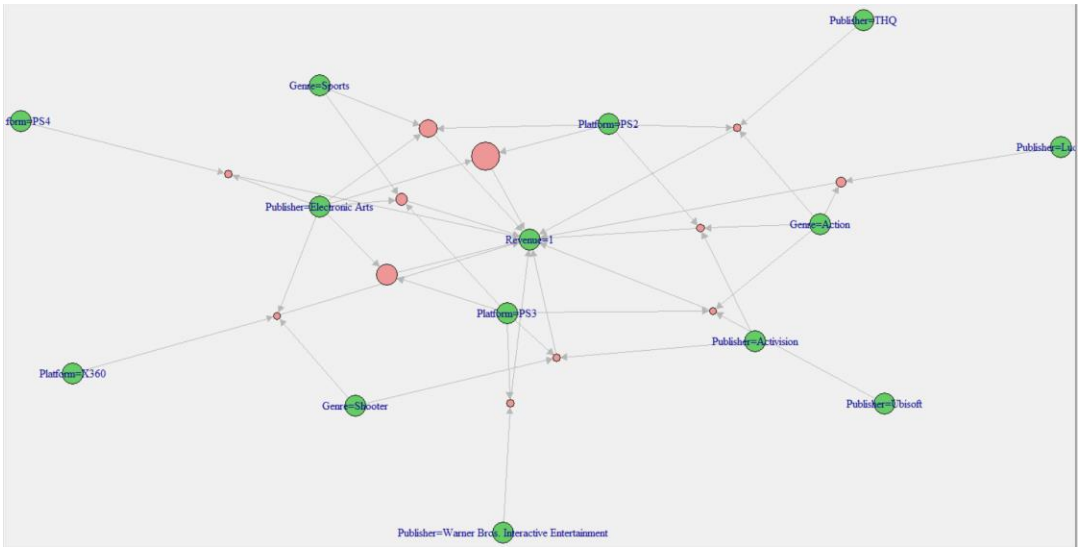


JP

```
     lhs                                                      rhs          support     confidence coverage    lift     count
[1]  {Platform=GB,Publisher=Nintendo}                      => {Revenue=1} 0.003675570 1.0000000  0.003675570 5.590411 60
[2]  {Platform=GB,Genre=Role-Playing}                      => {Revenue=1} 0.001286449 1.0000000  0.001286449 5.590411 21
[3]  {Platform=NES,Publisher=Nintendo}                     => {Revenue=1} 0.002756677 1.0000000  0.002756677 5.590411 45
[4]  {Platform=NES,Genre=Platform}                         => {Revenue=1} 0.001715266 1.0000000  0.001715266 5.590411 28
[5]  {Platform=SNES,Publisher=Nintendo}                    => {Revenue=1} 0.002205342 1.0000000  0.002205342 5.590411 36
[6]  {Platform=NES}                                        => {Revenue=1} 0.005819652 0.9693878  0.006003431 5.419276 95
[7]  {Genre=Role-Playing,Publisher=Enix Corporation}       => {Revenue=1} 0.001347709 0.9565217  0.001408968 5.347350 22
[8]  {Genre=Role-Playing,Publisher=SquareSoft}             => {Revenue=1} 0.002205342 0.9473684  0.002327861 5.296179 36
[9]  {Platform=SNES,Publisher=Hudson Soft}                 => {Revenue=1} 0.001041411 0.9444444  0.001102671 5.279833 17
[10] {Platform=GB}                                         => {Revenue=1} 0.005574614 0.9381443  0.005942171 5.244612 91
> plot(goodrules,method="graph",interactive=TRUE,shading=NA)
```



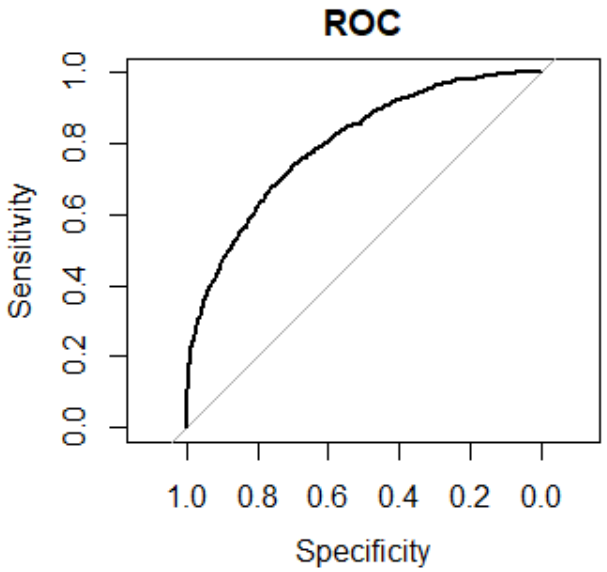
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{Platform=PS4,Publisher=Electronic Arts}	=>	{Revenue=1}	0.001347709	0.8461538	0.001592747	4.012962	22
[2]	{Platform=PS2,Genre=Sports,Publisher=Electronic Arts}	=>	{Revenue=1}	0.005145798	0.8316832	0.006187209	3.944334	84
[3]	{Platform=PS3,Genre=Action,Publisher=Ubisoft}	=>	{Revenue=1}	0.001041411	0.8095238	0.001286449	3.839241	17
[4]	{Platform=PS2,Genre=Action,Publisher=Activision}	=>	{Revenue=1}	0.001347709	0.7857143	0.001715266	3.726322	22
[5]	{Platform=X360,Genre=Shooter,Publisher=Electronic Arts}	=>	{Revenue=1}	0.001041411	0.7727273	0.001347709	3.664730	17
[6]	{Platform=PS2,Publisher=Electronic Arts}	=>	{Revenue=1}	0.009188924	0.7653061	0.012006861	3.629534	150
[7]	{Platform=PS3,Publisher=Warner Bros. Interactive Entertainment}	=>	{Revenue=1}	0.001347709	0.7586207	0.001776525	3.597828	22
[8]	{Platform=PS3,Publisher=Electronic Arts}	=>	{Revenue=1}	0.006493506	0.7517730	0.008637589	3.565352	106
[9]	{Platform=PS3,Genre=Sports,Publisher=Electronic Arts}	=>	{Revenue=1}	0.002940456	0.7500000	0.003920608	3.556944	48
[10]	{Platform=PS2,Genre=Action,Publisher=THQ}	=>	{Revenue=1}	0.001163930	0.7307692	0.001592747	3.465740	19



Other Areas

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4266 1237
         1  304  723

               Accuracy : 0.764
                 95% CI : (0.7535, 0.7743)
    No Information Rate : 0.6998
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3499

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9335
            Specificity : 0.3689
         Pos Pred Value : 0.7752
         Neg Pred Value : 0.7040
             Prevalence : 0.6998
         Detection Rate : 0.6533
   Detection Prevalence : 0.8427
      Balanced Accuracy : 0.6512

       'Positive' Class : 0
```

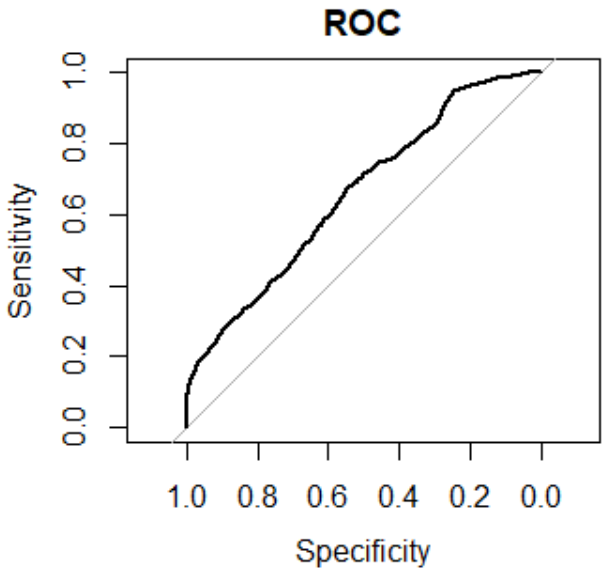
```
+   arrange(desc(Accuracy))
   k  Accuracy     Kappa  AccuracySD    KappaSD
1  9 0.7538287 0.2597847 0.007193806 0.02566645
2 13 0.7529098 0.2356891 0.003392384 0.01007369
3  7 0.7525014 0.2706327 0.005322695 0.01519909
4 11 0.7516842 0.2424620 0.006634558 0.01769081
5  5 0.7491313 0.2772117 0.011234572 0.02353896
> # results for best model
> confusionMatrix(knn)
Cross-Validated (3 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 67.4 19.0
         1  5.6  8.0

 Accuracy (average) : 0.7538
```



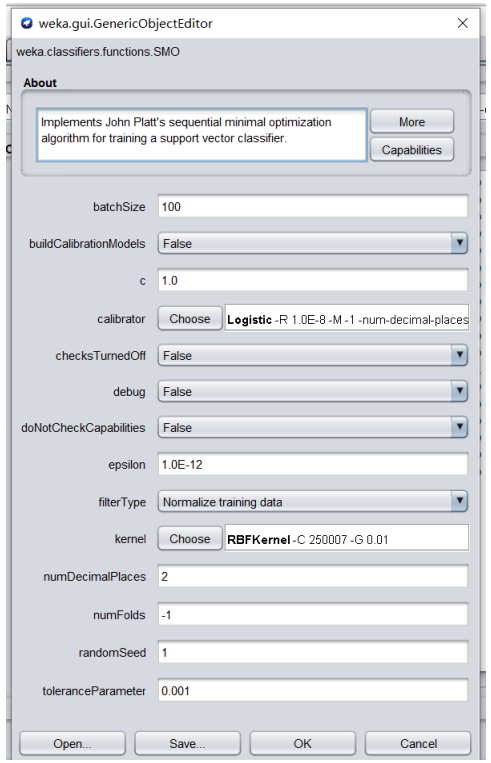
KNN

```
Confusion Matrix and Statistics

          Reference
Prediction  low high
      low  4357 1513
      high  213  447

               Accuracy : 0.7357
                 95% CI : (0.7248, 0.7463)
    No Information Rate : 0.6998
    P-Value [Acc > NIR] : 8.87e-11

                  Kappa : 0.2239

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9534
            Specificity : 0.2281
         Pos Pred Value : 0.7422
         Neg Pred Value : 0.6773
             Prevalence : 0.6998
         Detection Rate : 0.6672
   Detection Prevalence : 0.8989
      Balanced Accuracy : 0.5907

       'Positive' Class : low
```

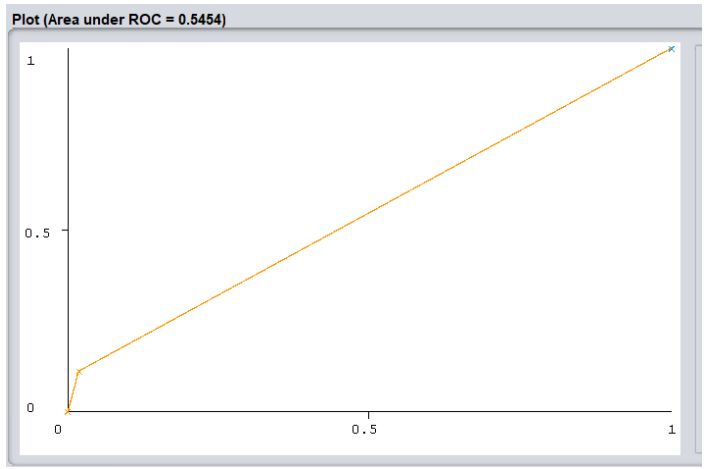
```
+   arrange(desc(Accuracy))
          C  Accuracy     Kappa  AccuracySD    KappaSD
1 0.2105263 0.7503573 0.1767374 0.002815791 0.02766899
2 0.3157895 0.7499489 0.1797110 0.002155321 0.02817966
3 0.4210526 0.7497447 0.1792915 0.001838205 0.02805097
4 0.1052632 0.7490300 0.1659823 0.002950048 0.02926081
5 0.5263158 0.7488259 0.1802332 0.003101134 0.03430902
6 0.6315789 0.7488259 0.1802332 0.003101134 0.03430902
> # results for best model
> confusionMatrix(svm)
Cross-Validated (3 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  low high
      low  70.4 22.4
      high  2.6  4.6

 Accuracy (average) : 0.7504
```



SVM with Linear kernel



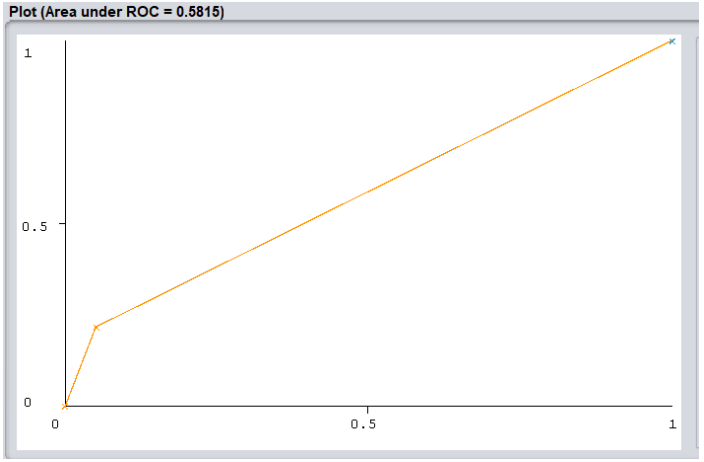

```
Time taken to build model: 2781.9 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       12494               76.5376 %
Incorrectly Classified Instances      3830               23.4624 %
Kappa statistic                          0.1257
Mean absolute error                      0.2346
Root mean squared error                  0.4844
Relative absolute error                 63.1124 %
Root relative squared error            112.3533 %
Total Number of Instances            16324

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
               0.111    0.020    0.642      0.111    0.189      0.194    0.545     0.291     hi
               0.980    0.889    0.771      0.980    0.863      0.194    0.545     0.771     lo
Weighted Avg.  0.765    0.675    0.739      0.765    0.697      0.194    0.545     0.652

=== Confusion Matrix ===

     a     b   <-- classified as
   447  3581 |     a = high
   249 12047 |     b = low
```



SVM with RBF kernel

batchSize	100
buildCalibrationModels	False
c	1.0
calibrator	Choose **Logistic** -R 1.0E-8 -M -1 -num-decimal-places
checksTurnedOff	False
debug	False
doNotCheckCapabilities	False
epsilon	1.0E-12
filterType	Normalize training data
kernel	Choose **RBFKernel** -C 250007 -G 0.01
numDecimalPlaces	2
numFolds	-1
randomSeed	1
toleranceParameter	0.001

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       12526               76.7336 %
Incorrectly Classified Instances      3798               23.2664 %
Kappa statistic                          0.2066
Mean absolute error                      0.2327
Root mean squared error                  0.4824
Relative absolute error                 62.5851 %
Root relative squared error            111.883  %
Total Number of Instances            16324

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
               0.214    0.052    0.577      0.214    0.313      0.243    0.581     0.318     hi
               0.948    0.786    0.787      0.948    0.860      0.243    0.581     0.785     lo
Weighted Avg.  0.767    0.604    0.735      0.767    0.725      0.243    0.581     0.670

=== Confusion Matrix ===

     a     b   <-- classified as
   864  3164 |     a = high
   634 11662 |     b = low
```



SVM with Polynomial kernel

	Accuracy	Rank	
KNN	75.38%	3	
SVM with Linear kernel	75.04%	4	
SVM with RBF kernel	76.73%	1	
SVM with Poly kernel	76.54%	2	
	Association Rules	Confidence	Support
Global Sales	Platform=NES,Publisher=Nintendo	1.0000	0.0028
	Platform=NES,Genre=Platform	1.0000	0.0017
NA Sales	Platform=2600,Publisher=Activision	0.9200	0.0014
	Platform=2600,Genre=Shooter	0.8636	0.0012
EU Sales	Platform=X360,Genre=Shooter,Publisher=Electronic Arts	0.7727	0.0010
	Platform=PS4,Publisher=Electronic Arts	0.7308	0.0012
JP Sales	Platform=GB,Publisher=Nintendo	1.0000	0.0037
	Platform=GB,Genre=Role-Playing	1.0000	0.0013
Other Area Sales	Platform=PS4,Publisher=Electronic Arts	0.8462	0.0013
	Platform=PS2,Genre=Sports,Publisher=Electronic Arts	0.8317	0.0051
	Platform=PS3,Genre=Action,Publisher=Ubisoft	0.8095	0.0010
	Platform=PS2,Genre=Action,Publisher=Activision	0.7857	0.0013

Conclusion

Final words

- Keep learning
- Data driven
- Business communication based on data science methodology
- Advanced data science technology (AI, Cloud, DW)
- Offer: I get a data analyst offer in Shanghai for production analysis.

Thank you for your listening

References

Applied Data Science Master's Degree. (n.d.). *ISchool | Syracuse University*. Retrieved February 18, 2022, from https://ischool.syr.edu/academics/applied-data-science-masters-degree/

Program: Applied Data Science, MS - Syracuse University—Acalog ACMS™. (n.d.). Retrieved February 18, 2022, from http://coursecatalog.syr.edu/preview_program.php?catoid=18&poid=9483&returnto=2322

ADS-Project-Portfolio-Milestone/README.md at main · wozhouwozhou/ADS-Project-Portfolio-Milestone. (n.d.). GitHub. Retrieved February 18, 2022, from https://github.com/wozhouwozhou/ADS-Project-Portfolio-Milestone

Hotel booking demand. (n.d.). Retrieved February 19, 2022, from https://kaggle.com/jessemostipak/hotel-booking-demand

wozhouwozhou. (2022). ADS-Project-Portfolio-Milestone [Jupyter Notebook]. https://github.com/wozhouwozhou/ADS-Project-Portfolio-Milestone