Data Science for Business

E – Above and Smart Real Estate Projects

Final Report

Master in Business Consulting
Winter Semester, February 2018

Submitted to

Professor Dr. Holger Ziekow

Prepared by

Juan Carlos Aguila
Ivan Rojas
Lei Chenlong

Fakultät Wirtschaftsinformatik
Hochschule Furtwangen University

INTRODUCTION

This report is based on data from two business cases E- Above related to taxi operations in NYC and Smart Real Estate about investments in housing in the United Kingdom.

Sample of 30,000 observations for both data sets have been explore, prepared and then analyzed with the help of programming language R. Patterns and tendencies led us to formulate clear statements and the report displays graphs and maps generated by the software which helps the reader to have a better interpretation of the main findings.

The main focus resides in identifying a cost structure and the interaction between different variables that help to explain how they behave and its impact in the final price. Data is transform into valuable information to facilitate the decision making process and ultimately made more profit for the business.

The final part of the project includes annexes with the coding done during the exercise and main conclusions.

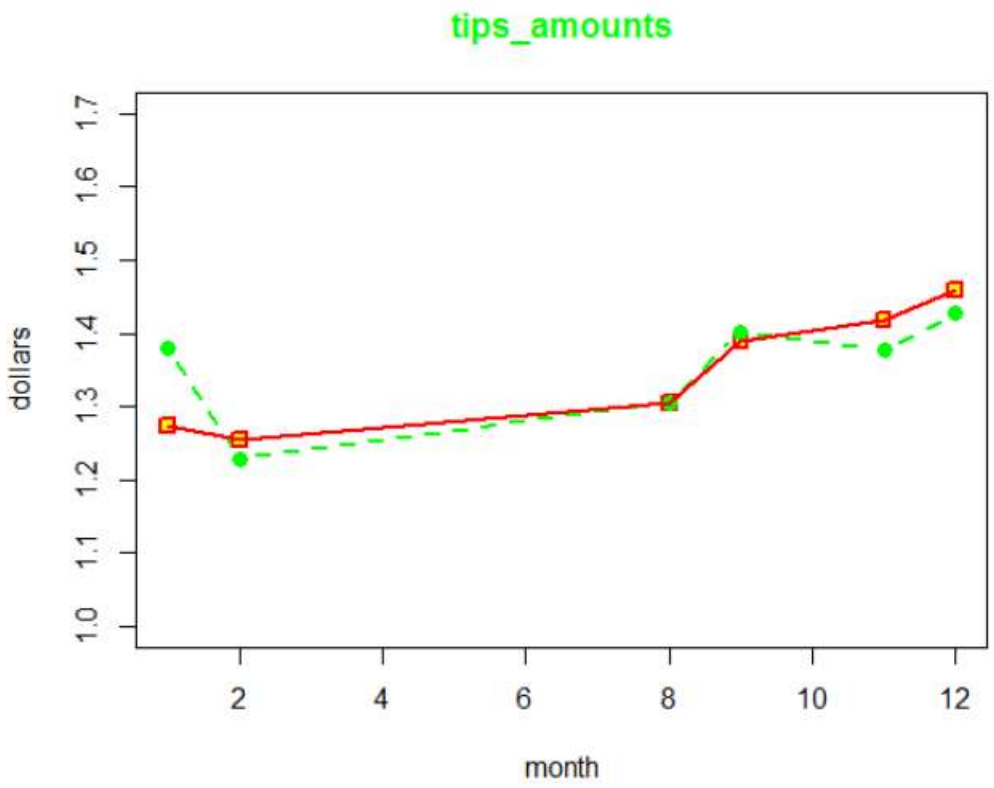
E – ABOVE PROJECT

NYC taxi vendors and drives can get useful suggestions from the conclusions.

We come up with the following questions:
1. Is there any difference for tips received between the drivers from two vendors?
2. Which vendor covers larger market?
3. When is better to earn more money?
4. Is there any relation between the tips and distances?
5. Payment difference exists or not?

According to these questions, we used the data and got the conclusions:

1. By drawing data of 6 months from the file randomly, we visualize it and find out that there is no huge difference between KMT and VTS.



2. We get 3000 data entries and find that the number for each is around 1500.

newdataVTS88	1516 obs. of 11 variables
newdataVTS99	1503 obs. of 11 variables

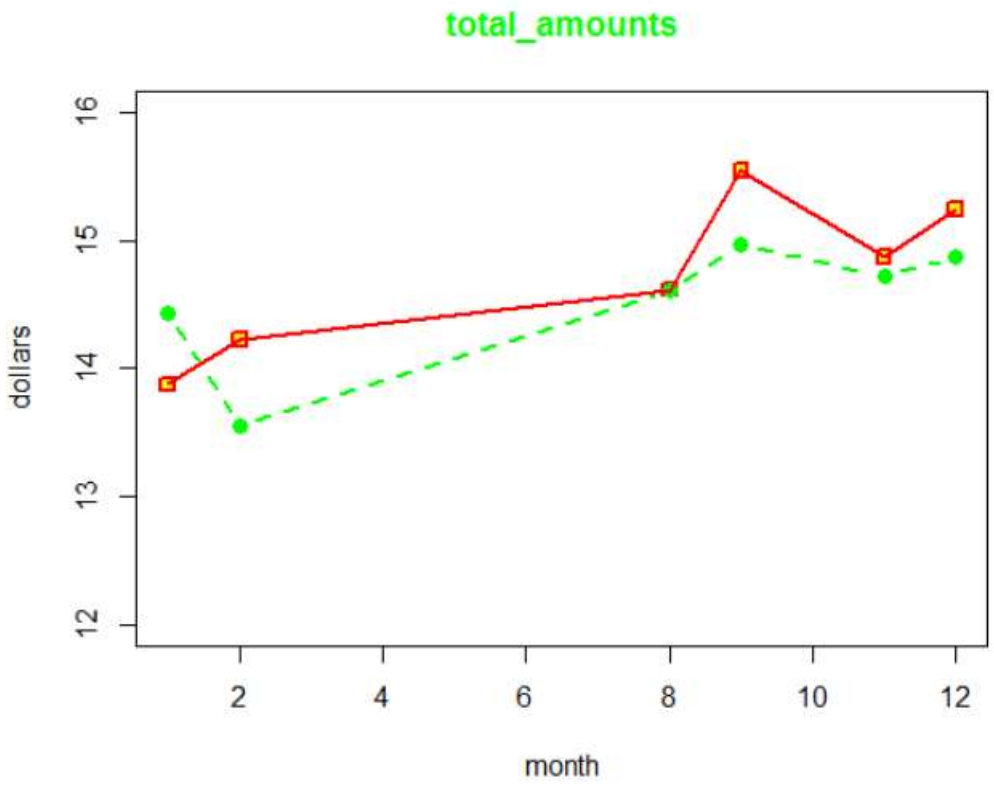
newdataVTS11	1472 obs. of 11 variables
newdataVTS1111	1511 obs. of 11 variables
newdataVTS1212	1561 obs. of 11 variables
newdataVTS22	1420 obs. of 11 variables

newdataCMT88	1484 obs. of 11 variables
newdataCMT99	1497 obs. of 11 variables

newdataCMT11	1528 obs. of 11 variables
newdataCMT1111	1489 obs. of 11 variables
newdataCMT1212	1439 obs. of 11 variables
newdataCMT22	1580 obs. of 11 variables

Then we can conclude that the two vendors, VTS and CMT, have almost the same shares of the taxi-running market of New York City.
This is a good phenomenon since there is no monopoly and the competitions between different vendors can improve the taxi service.

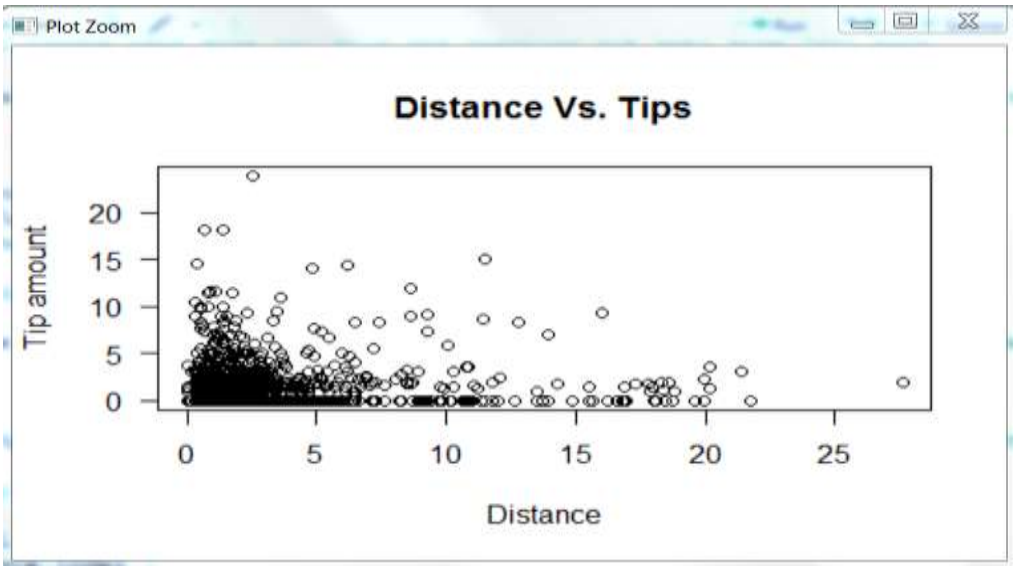
3. Just like the way for the question 1, we get the similar chart for the total fare amounts of 6 months.

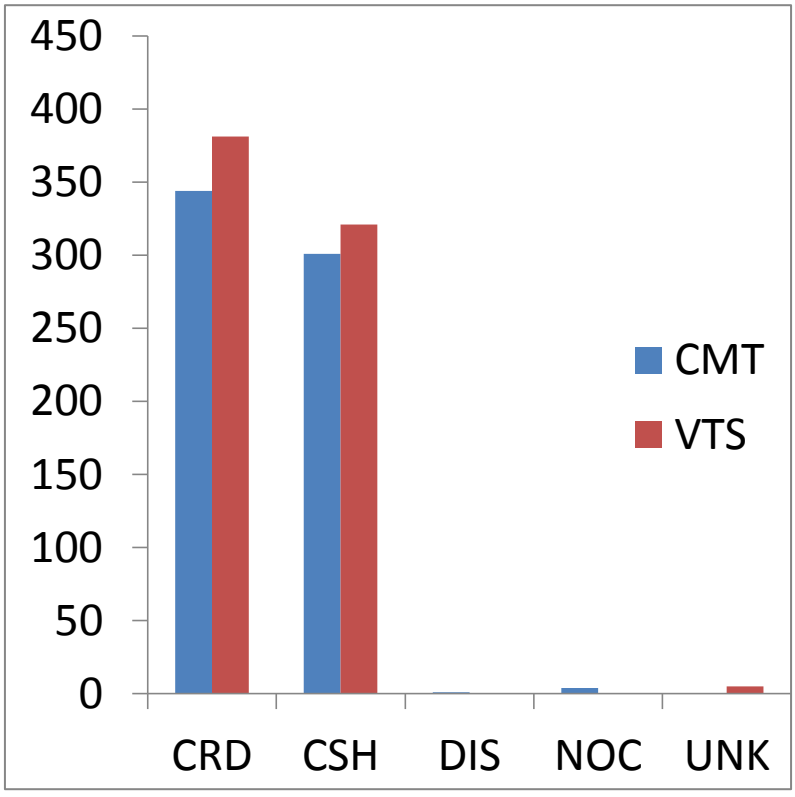


From the twisted lines, we can see that differences from the two are not big.

And we would recommend the taxi drivers to work more in August, September, November and December.

4. Dots in the chart gather together and overlap in the left bottom part, which means tips won't be high but people will give so it won't be a bad choice to take many short-distance trips.



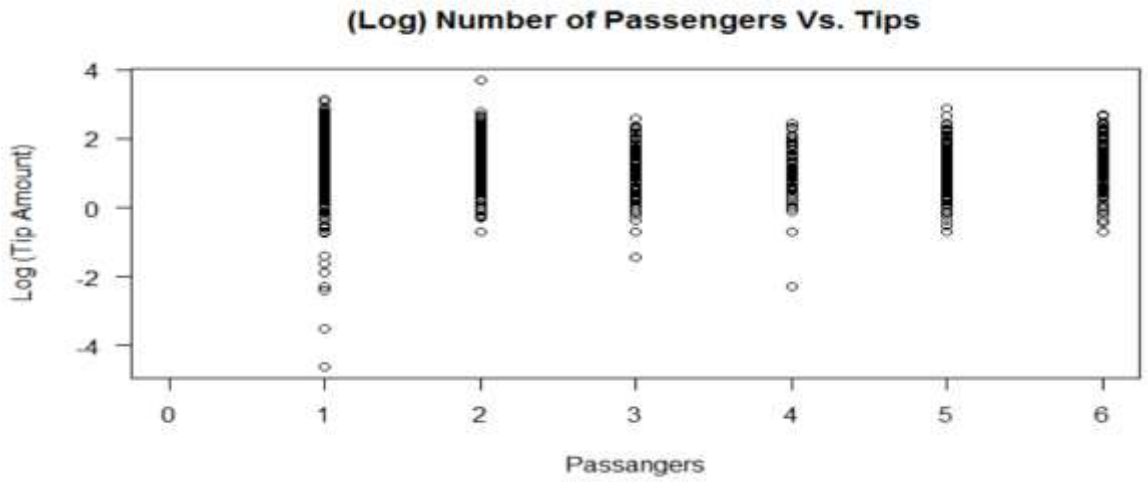


CRD = Cards debit / credit
CSH = Cash
DIS = Dispute fare
NOC = No charge
UNK = Unknown

Paying by cards and cash almost share half of the ways. And, the idea of society without cash is getting popular in recent years for reducing the possibility of getting fake currencies. Besides, Paying by smart phone apps seems not to appear in NYC taxi. If the NYC wants to develop the cashless society, they can begin to do some investigations to achieve that.

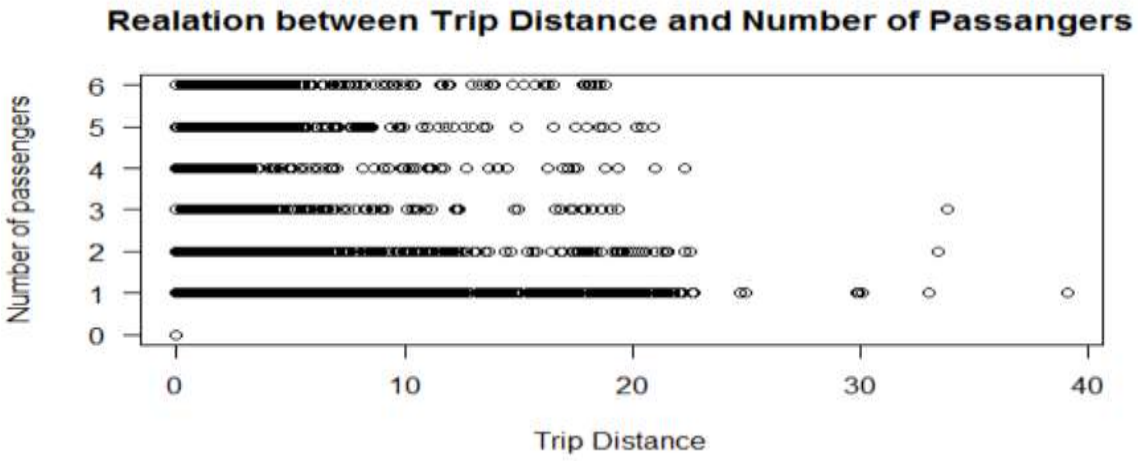
Tips Vs Passengers

For the project we have a maximum capacity of six passengers for trip that was distributed as follows:



According to the graph, trips with only one passenger contributed with the highest level of tips whereas trips taking three or four passengers count for the lowest tips amounts. Most of the tips are in the range of exp (0) to exp (3) equivalent from 1 to 20 dollars.

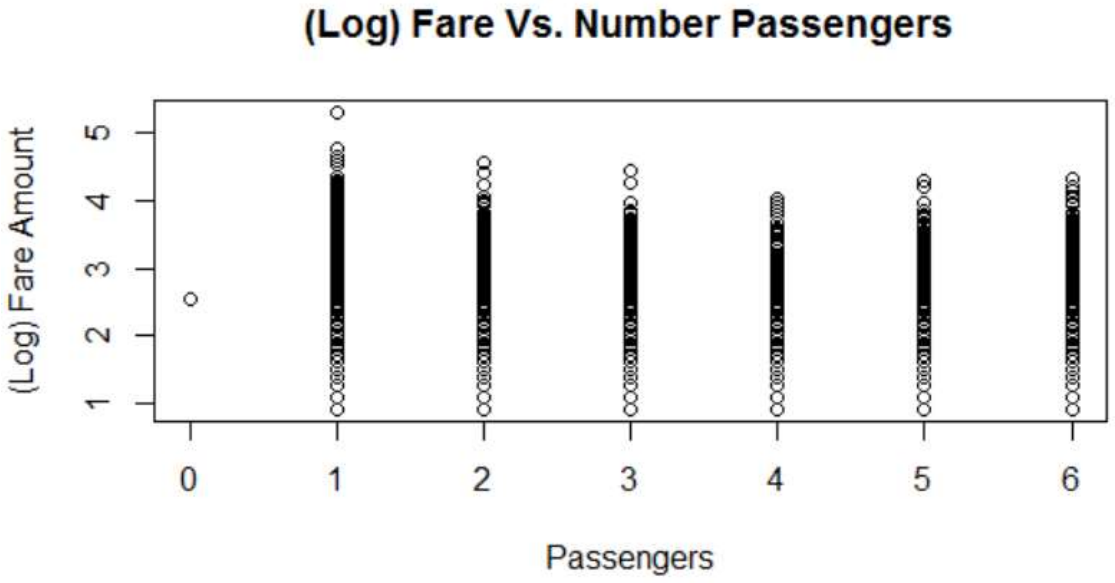
Passengers Vs. Distance



Based on the observation of the above chart we can conclude that longer distances are associated with one or two passengers. Conversely, four passengers traveled the shortest distances.

Most of trips were in the range between 0.1 and 22 kilometers.

Fare Vs. Passengers

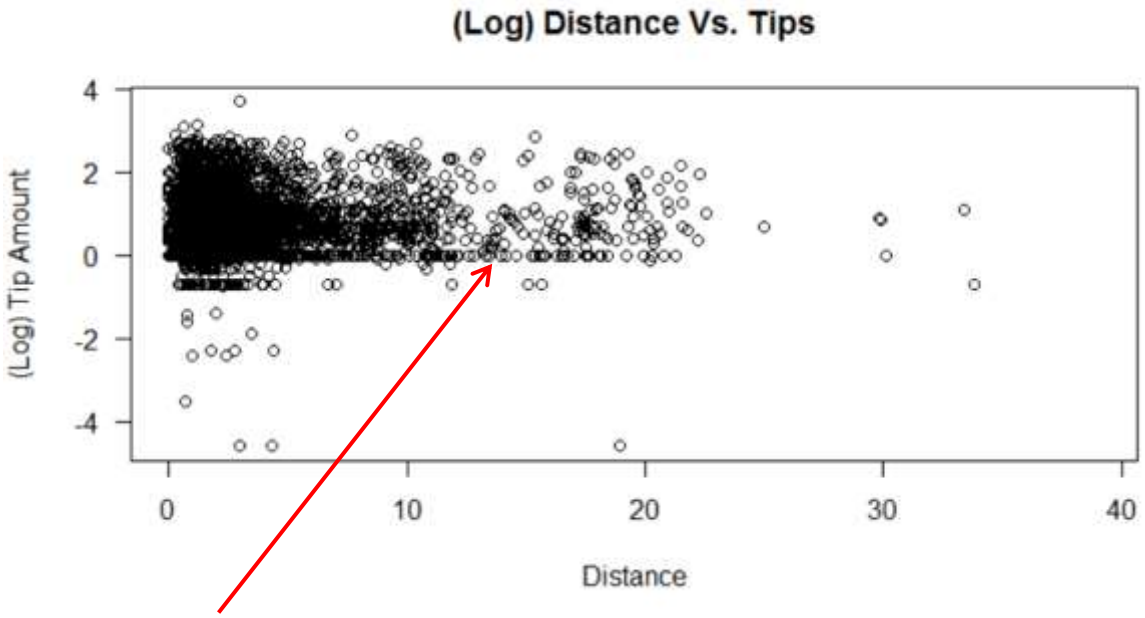


The highest fare amounts were collected in one passenger trips. On the other hand, trips with four passengers reported lower fare amounts.

The average fare amount was between exp(3) to exp(4) which equals 20 to 55 dollars.

Distance Vs. Tips

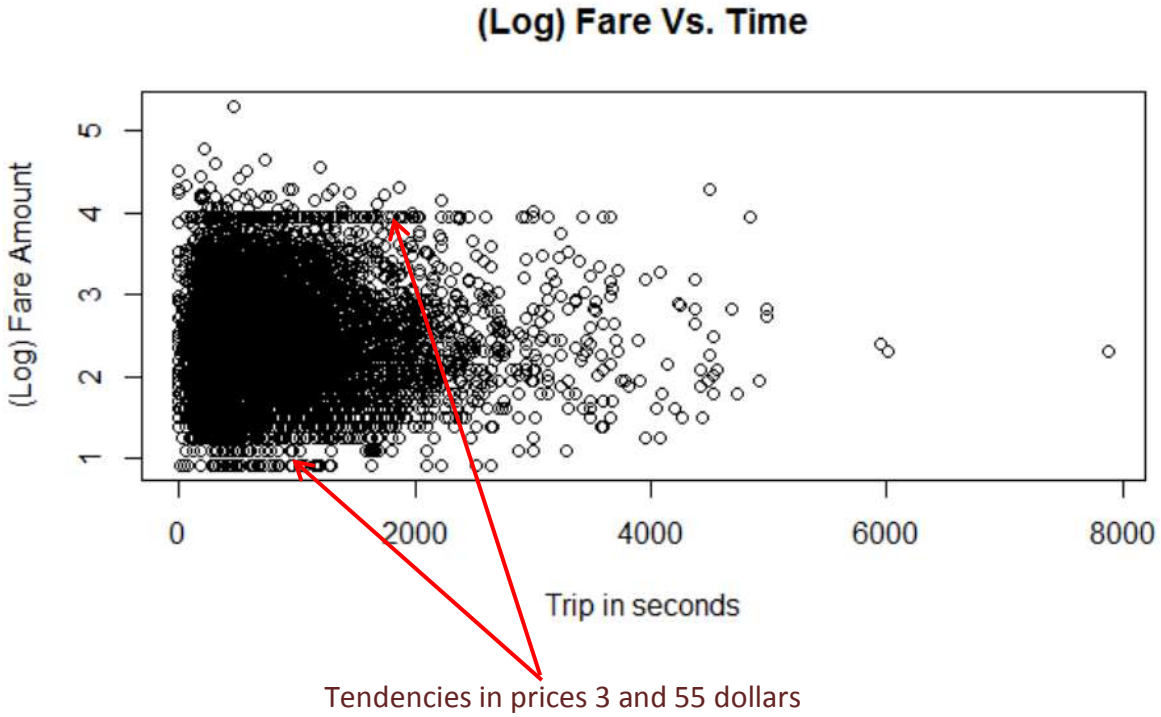
Most of the tips are associated with trips under 5 kilometers and they count exp(0) to exp(2) or 1 to 7 dollars. Longer distances did not necessarily represent higher tips.



It can also be noted that 1 dolar or exp(0) is the most recurrent tip amount disregarding the distance.

Fare Vs. Trip Duration

The vast mayority of trips had a duration of 30 minutes or below.
Fares were within the range of exp(1) to exp(4) or 3 to 55 dollars with well define tendencies along these amounts



Maps for pickups and drop off in NYC area

The three airports; 1. Newark Liberty International Airport, 2. LaGuardia, 3. John F Kennedy Airport, are locations away from the city center but with high density of taxi services. The first airport, located in New Jersey, present the lowest number of trips compared with the other two.

Manhattan is the area with the highest concentration where taxis are hailed and people dropped off, followed by the west areas of Brooklyn and Queens Districts.

Pickup's

Airports

NYC airports and their connection with Manhattan is an important corridor which can bring additional opportunities like shuttle bus services for E - Above focusing on travelers.

Drop off

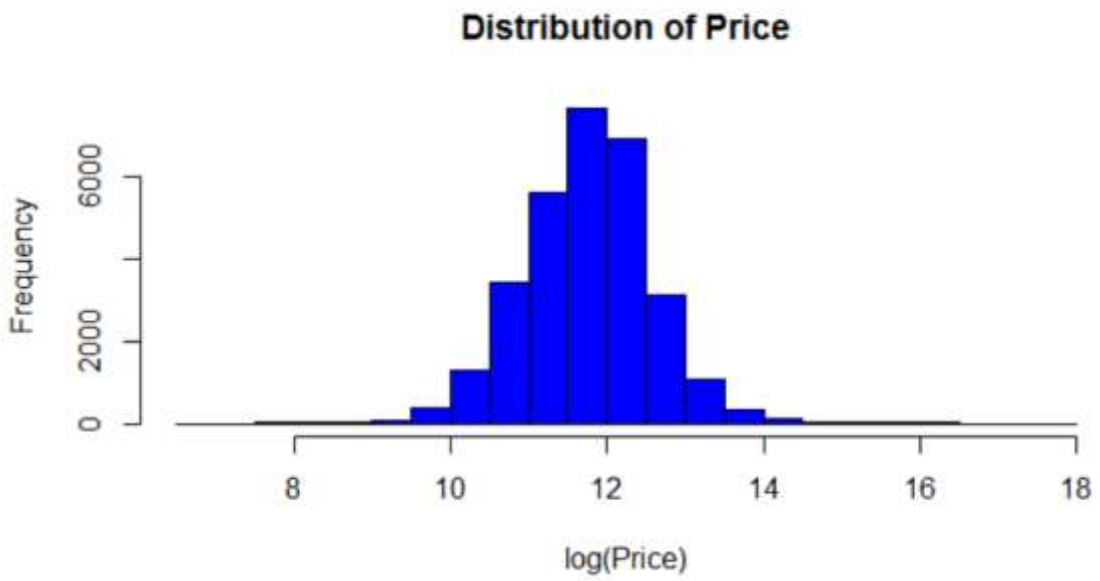

Pelham Bay Park
BRONX
East Rutherford
North Bergen
Secaucus
COLLEGE POINT
FLUSHING
JACKSON HEIGHTS
ELMHURST
FOREST HILLS
Kearny
Harrison
Newark
Jersey City
York
BUSHWICK
BROOKLYN
PARK SLOPE
EAST NEW YORK
Bayonne
NEW JERSEY
NEW YORK
ST. GEORGE
BAY RIDGE
Upper Bay

UNITED KINGDOM PROPERTY HOUSES PROJECT

A random sample of 3000 properties with 11 variables was selected from the original data set for a total period of 22 years from 1995 to 2017.

Price analysis

Prices were ranging between 750 up to £ 4,875,000, with a mean of £ 181,895.

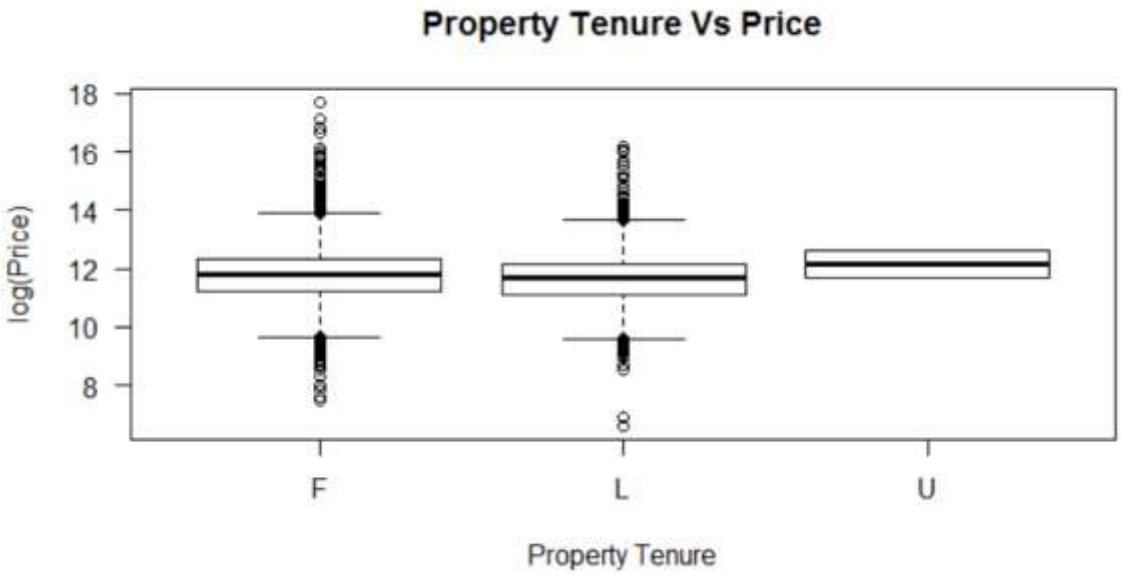


Based on the above graph it is observed that log(12) equivalent to £ 162,755 is very close the mean of prices.

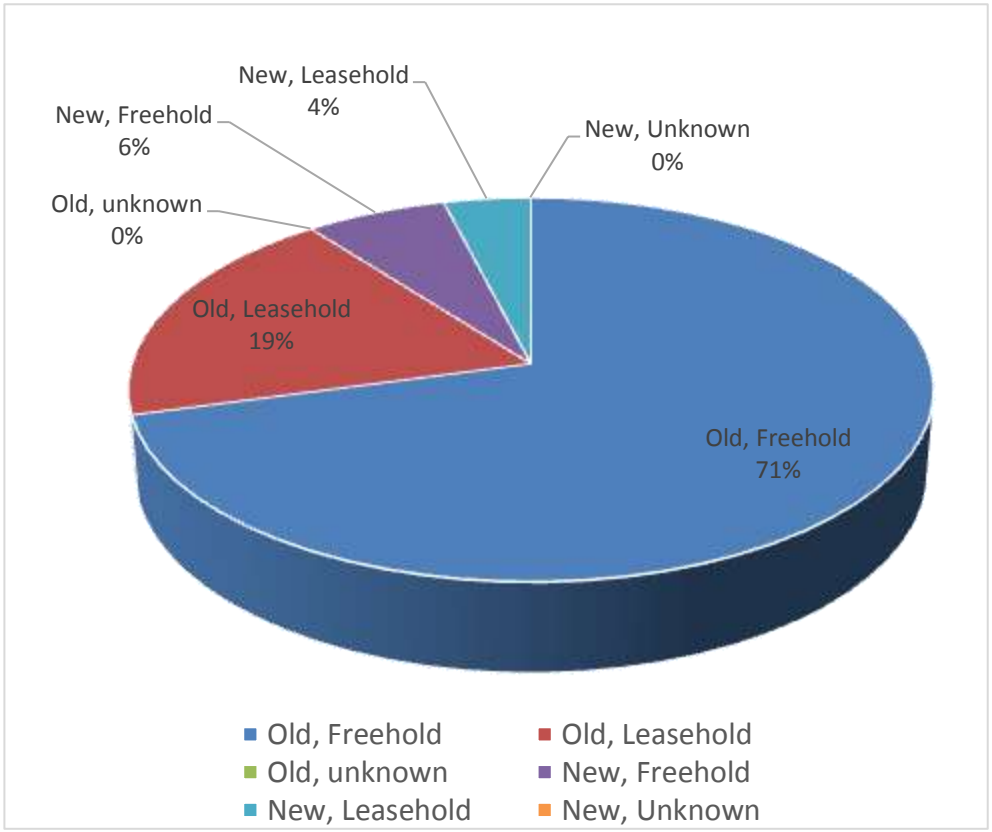
Property prices explained by Tenure

F=Freehold, L=leasehold U= unidentified

Unidentified tenure (U) price was slightly higher than the other two categories, however, price variations is grater for Freehold properties which could mean that they are more actively traded enticing the supply and demand in the real state UK market.



Relationship between (N=Old Y=New) and Duration (F=Freehold, L=leasehold U= unidentified)



In the above graph we can see than most of the houses are old and freehold (71%). The same pattern is seen in the type of properties which are new, most of them are freehold, but in this case the difference is not too big as in the first case. As a conclusion, we can say that freehold is the most popular type of house duration in the United Kingdom.

Prices explained by age of the property

N = Old; Y = New

Mean on prices for old or new properties in the UK are basically the same, around £ 163,000 or exp(12). Prices with higher variations correspond mostly to the Old properties. It can also be noticed that ranges for New (Y) properties prices tend to be lower.
Based on this observation, the company “Real State” might be able to find more opportunities in the market for old properties as prices present more variability ranging from £ 14,000 up to £ 1,200,000.

Prices explained by Property Types

D = Detached
A stand-alone **house, f**ree-standing residential building for a single family

S = Semi Detached
Two different properties divided by a wall

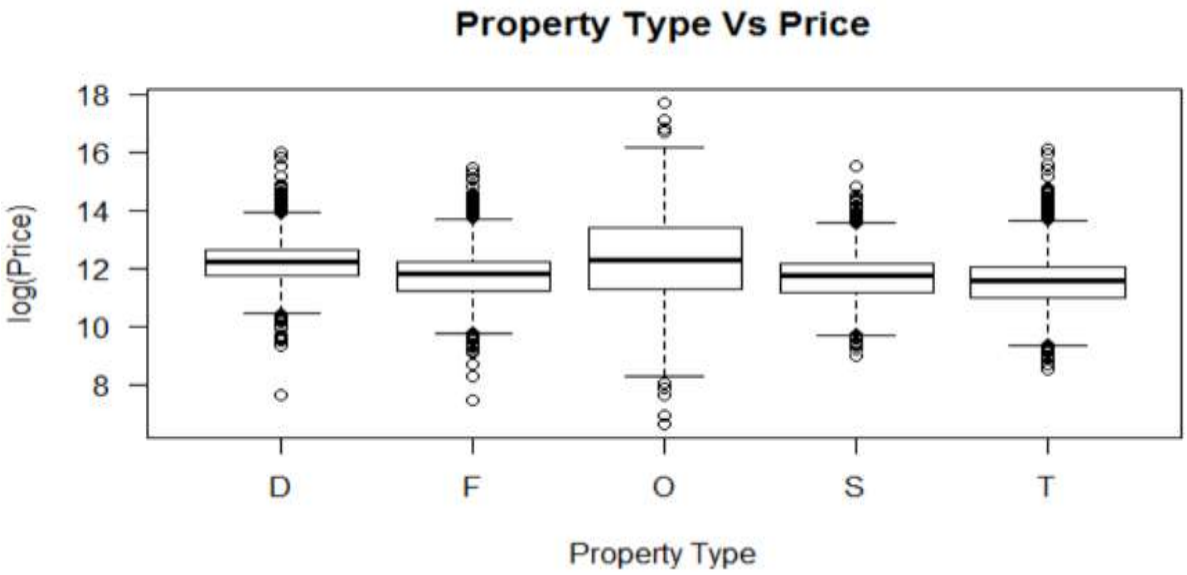
F = Flats
Building split up in multiple living areas for different residents

T = Terrace
Property situated in a long row of houses

O = Other
This category of properties is frequently located in rural areas and they are referred as Bungalows and cottages.

Detached properties are normally more expensive within the categories found in urban locations. Its minimum average price is also the highest with exp(10.5) or about £ 36,000.

Terrace type properties show the lowest mean price within urban areas categories, £ 120,500 and £ 13,000 for the lowest price.

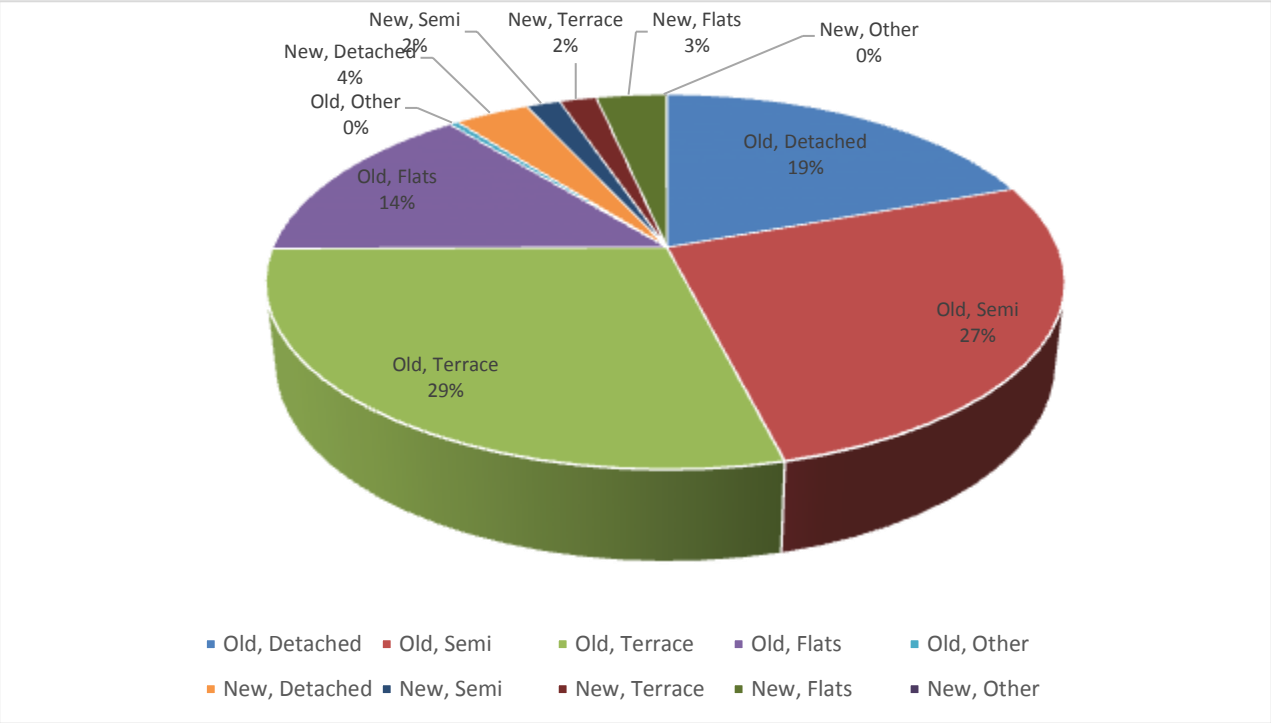


Finally Other (O) type of properties located mainly in rural areas present the highest variability of prices ranging between £ 3640 and £ 8,890,000 they also represent the smaller number counting 127 units analyzed.

Relationship between Property Type (D=Detached, S=Semi, T=Terrace, F=Flats O= Other) and (N=Old Y=New)

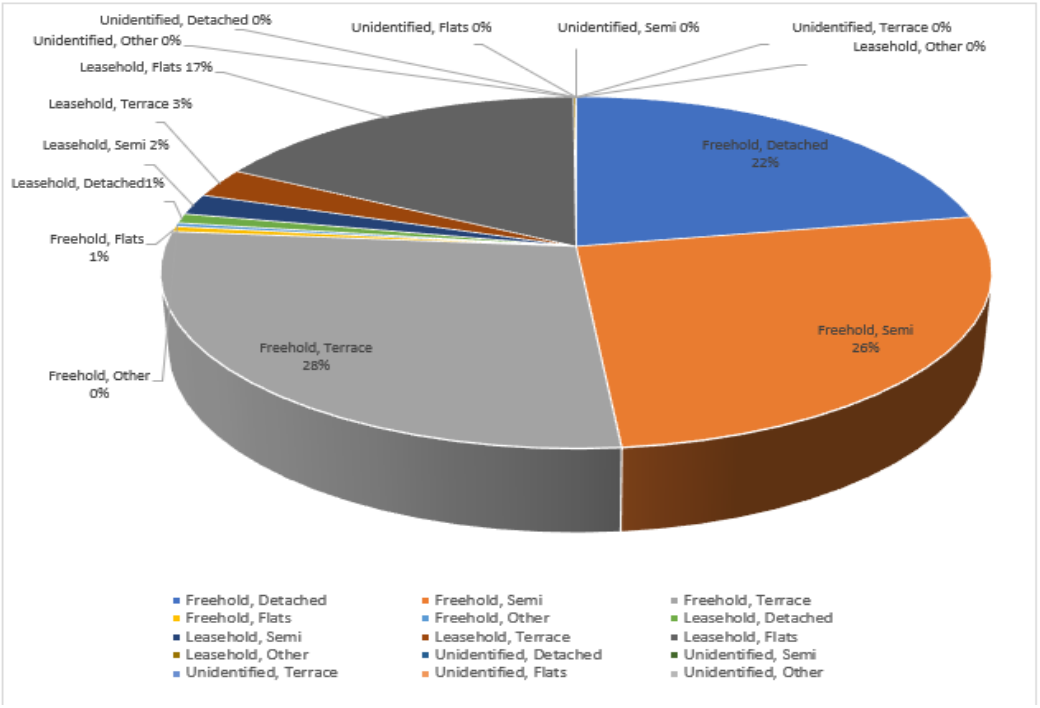
In this chart we can clearly see that among the property types terrace and semi-detached are the most popular when we take into consideration only the old houses, they are so popular that if we combine both of then they count more than 50% of the properties. On the other hand, when we consider only new houses, we

can see that detached and flats are the most popular property types, this means that is opposite scenario compared to the old houses.



It is also relevant to point out that the property type other is popular neither in old nor in new houses.

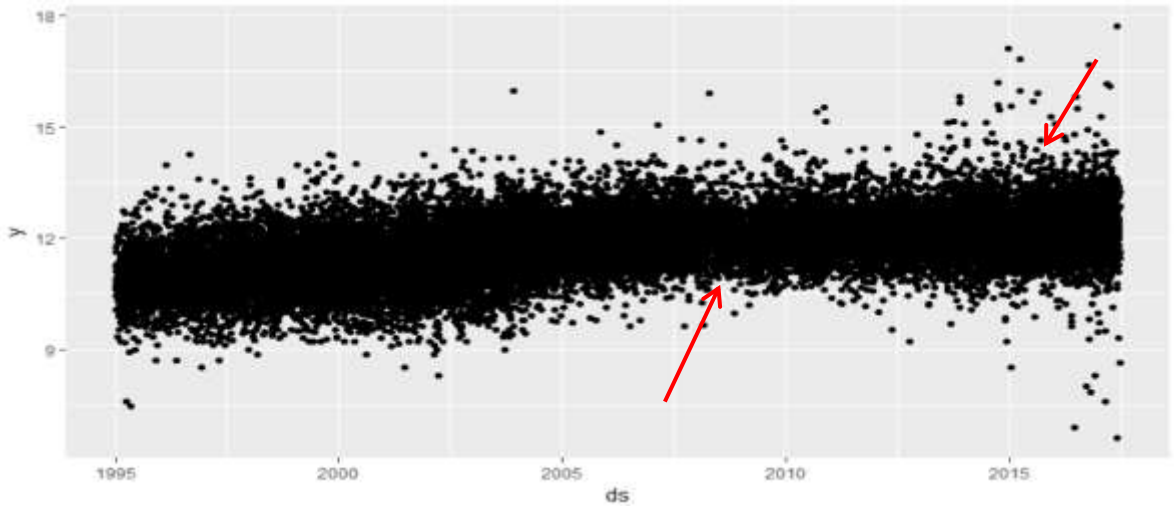
Relationship between Property Type (D=Detached, S=Semi, T=Terrace, F=Flats O= Other) and Duration (F=Freehold, L=leasehold U= unidentified)



In this diagram, it is perceived than in freehold houses there are just a few properties which fall under the category of flat (1%), on the contrary, if we take into consideration just properties which have leasehold duration the most popular kind of property type is flat, almost six times more popular than terrace which is the second one most common within leasehold properties. Also, is relevant to mention that properties with unidentified duration are not popular at all in the United Kingdom.

House prices predictive analysis

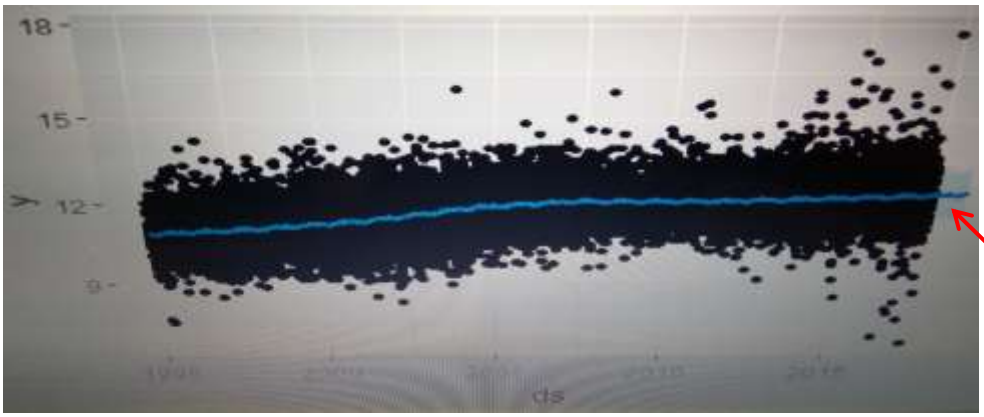
The chart below display two variables; y = log(price) and ds = dates, starting in 1995 until 2017.



It is noticed that dispersion of prices are shrinking by 2008 and 2009 presumably because of the world financial crisis. The dispersion becomes wider in 2015 and 2016 as a result of a general recovery of the economy.

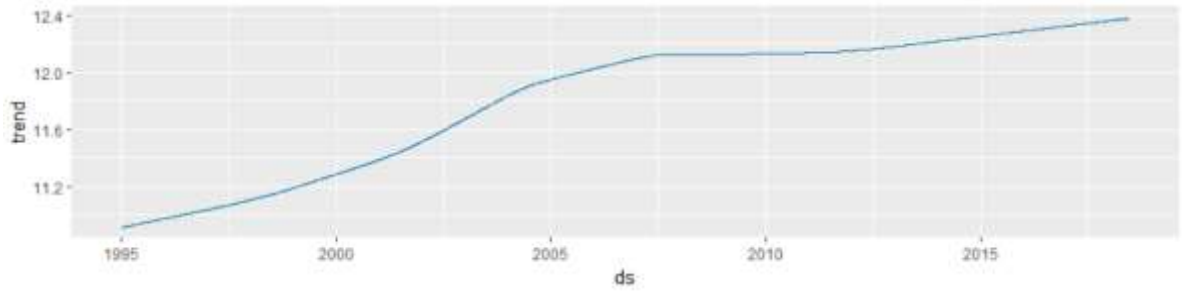
In order to do the prediction, hours and seconds were removed from the original data set and a package called Prophet developed by Facebook was utilized, this application is suitable for our purpose and additionally enable further analysis by displaying different components; trends, multiple seasonality and intervals.

The blue line illustrates the general trend of prices for the whole sample of properties and it was enquiry to predict 365 days which has been point out by the arrow.



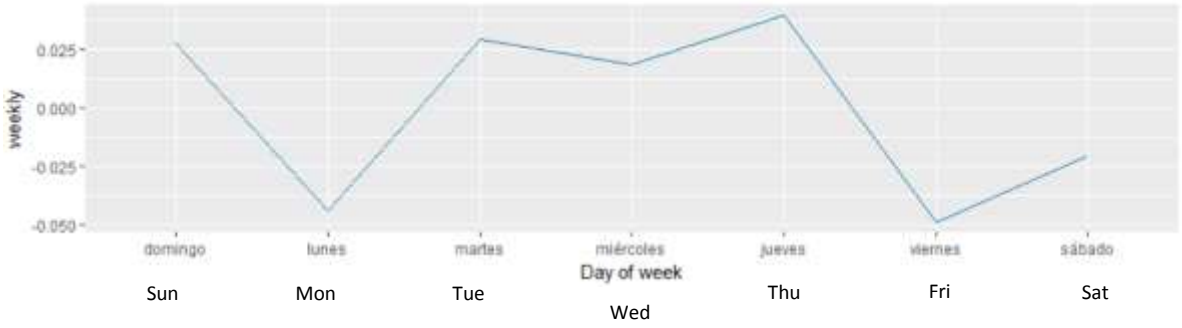
Between 1995 and 2007 there was a steady uptrend in prices which plateau for three years from 2008 until 2011. Afterwards prices have continuously increased until 2018.

General Trend



Important political events occurred in 2016 like the “Brexit” apparently did not impact on prices and the general uptrend has continued.

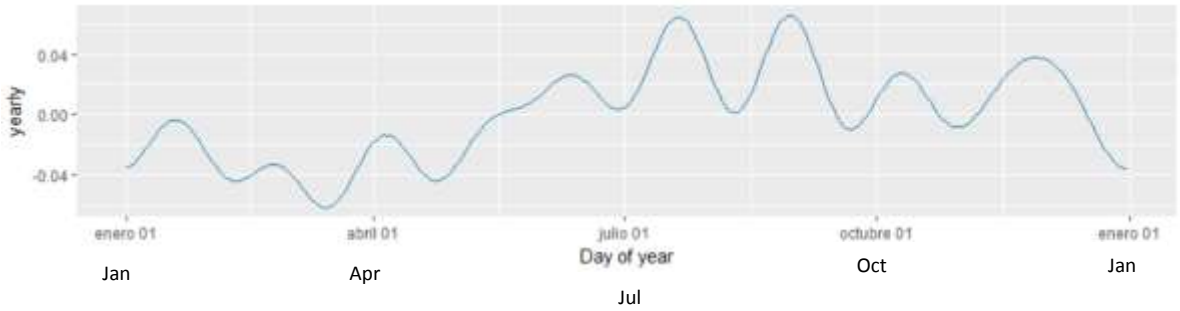
Daily Seasonality



According to the daily seasonality chart shown above, property prices tended to be lower on Monday and Fridays. On the hand, prices are higher from Tuesday to Thursday. Sundays

Monthly Seasonality

Prices are cheaper from January to March when lowest prices are reached. From end of March throughout August an increase in prices is noticed Prices soar by end of July and beginning of September. By the end of September prices would decline for the rest of the year.



Our recommendation for “Smart Real State” is to secure financial resources for purchases made in March each year as better deals can be found in the market. Following this statements, the company should avoid buying properties during July and September unless they have properties available to be sold to.

Maps for property location in United Kingdom

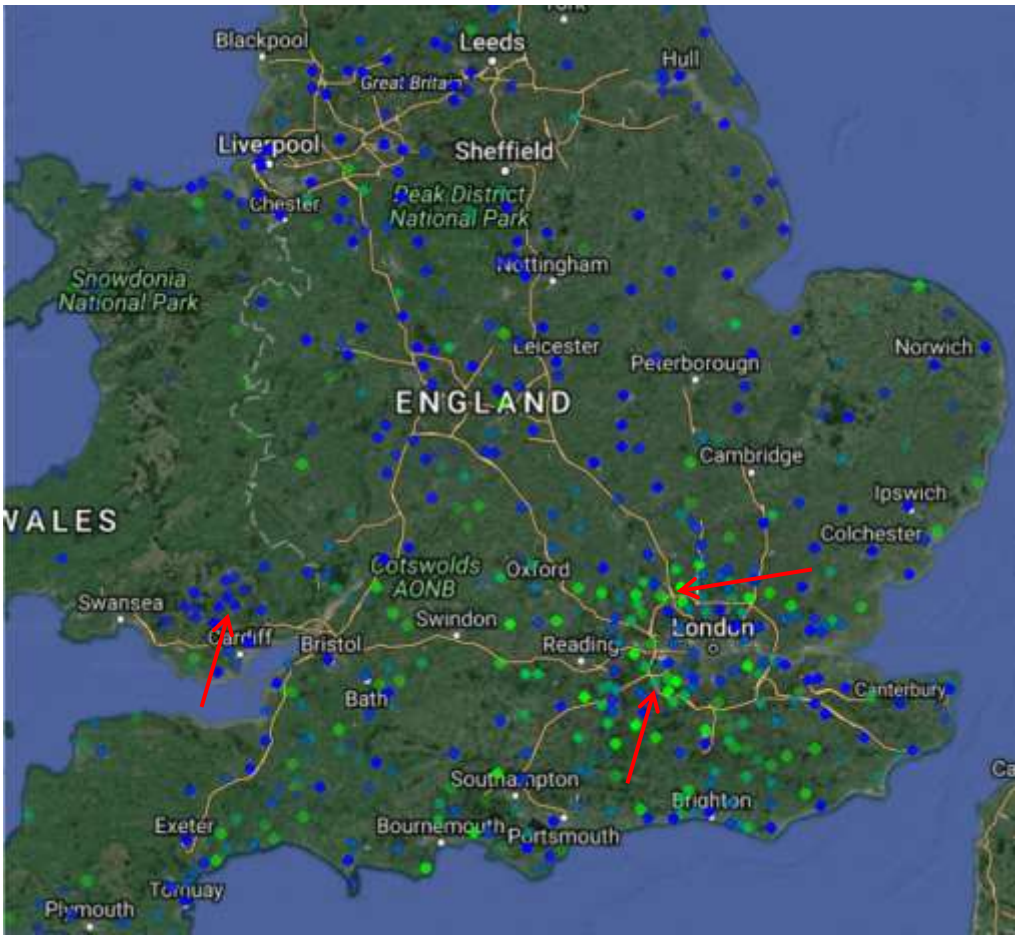
For this analysis, expensive properties are considered those that worth more than £ 210,000, shown in green color in the graph and cheaper properties are below £ 75,000 identified with blue signs.

It can be noticed that there is a significant density of expensive properties located in west and north of London. The situation becomes more disperse towards south of the city.

In the north of England prices are remarkably cheaper and lots of properties under £ 75,000 are found especially in rural areas such as Leeds and Liverpool.

Another important cluster where cheap properties are closely located is near Cardiff.

Depending on preferences and factors aside price, “Smart Real State” company could benefit of buying cheap properties in areas where expensive ones are located mainly in north and western neighborhoods of London. By doing so, the company might find better deals and after renovations they might gain more value if they are intended to be sold again following the general uptrend in prices.



CONCLUSIONS

Evidence has shown that there is no monopoly and the competition between different vendors can improve the taxi service.

Taxi services tend to have more dynamic activity during; August, September, November and December.

Paying by cards and cash almost share half of the ways of payment which in turn might generate an opportunity for companies like E – Above by encouraging alternative ways such as cashless payment through smart phone apps.

The three airports in NYC are locations away from the city center but with high density of services particularly LaGuardia and John F Kennedy. Their connection with Manhattan is an important corridor which can bring additional opportunities like shuttle services focusing on travelers.

For the United Kingdom property houses project prices were ranging between £ 750 up and £ 4,875,000, with a mean of 181,895 pounds.

Smart Real State could find more opportunities in the market for old properties as prices present more variability ranging from £ 14,000 up to £ 1,200,000

Cheaper properties are frequently located in rural areas and they fall in “Other” category and are described as Bungalows and cottages.

Detached properties are normally more expensive within the categories found in urban locations. Its minimum average price is also the highest with exp(10.5) or about £ 36,000.

It is noticed that dispersion of prices are shrinking by 2008 and 2009 presumably because of the world financial crisis. The dispersion becomes wider in 2015 and 2016 as a result of a general recovery of the economy.

Important political events occurred in 2016 like the “Brexit” apparently did not impact on prices and the general uptrend has continued.

According to the daily seasonality chart shown above, property prices tended to be lower on Monday and Fridays. On the hand, prices are higher from Tuesday to Thursday and Sundays.

Prices are cheaper from January to March when lowest prices are reached. From end of March throughout August an increase in prices is noticed Prices soar by end of July and beginning of September. By the end of September prices would decline for the rest of the year.

Smart Real State should secure financial resources for purchases made in March each year as better deals can be found in the market.

The company should avoid buying properties during July and September

Smart Real State could benefit of buying cheap properties in areas where expensive ones are located mainly in north and western neighborhoods of London.

ANNEX SCRIPS

E – Above

```
library(readr)
library(data.table)
library(rworldmap)
library(ggplot2)
library(ggmap)


save(taxi_trip,taxi_fare,merged1_clean,file = "NY.Rdata")
load("NY.Rdata")

# Selecting 30000 random rows from trip file

Taxi_Data <- fread("R:/Data_Science/Taxi_trip_Data/trip_Data_12.csv")
index_taxi_trip <- sample(1:nrow(Taxi_Data), 30000)
index_taxi <- Taxi_Data[index_taxi_trip,]
taxi_trip <- index_taxi
head(taxi_trip)

# Selecing 30000 random rows from fares file

Taxi1 <- fread("R:/Data_Science/Taxi_trip_fare/trip_fare_12.csv")
index_taxi_1 <- sample(1:nrow(Taxi1), 30000)
index_taxi_fare <- Taxi1[index_taxi_1,]
taxi_fare <- index_taxi_fare
summary(taxi_fare)
head(taxi_fare)
fix(taxi_fare)

medallion <- taxi_fare$medallion

payment_type <- taxi_fare$payment_type
class(payment_type)

fare_amount <- taxi_fare$fare_amount
class(fare_amount)

tip_amount <- taxi_fare$tip_amount

taxi_fare_cleaned <- data.frame(medallion, payment_type,fare_amount,
tip_amount)
head(taxi_fare_cleaned)
```

```
# joining both data frames

merged1 <- merge(x = taxi_trip, y = taxi_fare_cleaned, by = "medallion", all =
FALSE)
fix(merged1)

# Elimination of redundancies according to "medallion" as a taxi identification

merged1_clean <- merged1[!duplicated(merged1$medallion),]
str(merged1_clean)
fix(merged1_clean)


# Mapping locations in New York city where taxis pick up and drop off people

map_ny <- get_map("New York City", zoom = 15, maptype = "hybrid", source =
"google")
class(map_ny)

ggmap(map_ny) + geom_point(data = merged1_clean, aes(x = pickup_longitude, y
= pickup_latitude), color = "green", size = 2, alpha = 0.5)

ggmap(map_ny) + geom_point(data = merged1_clean, aes(x = dropoff_longitude,
y = dropoff_latitude), color = "red", size = 3, alpha = 0.5)



plot(log(merged1_clean$fare_amount)~merged1_clean$passenger_count, ylab =
"(Log) Fare Amount", xlab= "Passengers", main = "(Log) Fare Vs. Number
Passengers")
exp(1)

plot(log(merged1_clean$fare_amount)~merged1_clean$trip_time_in_secs, ylab =
"(Log) Fare Amount", xlab= "Trip in seconds", main = "(Log) Fare Vs. Time")
exp(3.91)

plot(log(merged1_clean$fare_amount), merged1_clean$trip_distance, ylab =
"(Log) Fare Amount", xlab = "Trip Distance", main = "(Log) Fare Vs. Distance")


# Using Plot to explain relation between distance and Tip amount variables

plot (merged1_clean$trip_distance,log(merged1_clean$tip_amount), xlab =
"Distance", ylab = "(Log) Tip Amount",
las = 1, main = "(Log) Distance Vs. Tips")
```

```
# Number of passengers Vs. amount of tip

plot (log(merged1_clean$tip_amount) ~ merged1_clean$passenger_count, xlab =
"Passangers", ylab = "Log (Tip Amount)",
     las = 1, main = "(Log) Number of Passengers Vs. Tips")

# Example for exp(2)= US$ 7.39

# Relation Trip Distance and Passangers number

plot (merged1_clean$passenger_count ~ merged1_clean$trip_distance , xlab =
"Trip Distance", ylab = "Number of passengers",
     las = 1, main = "Realation between Trip Distance and Number of Passangers")


# Payment type

table(merged1_clean$payment_type)


head(merged1_clean)

# How fare amount can be explained by other 4 different variables; Vendor,
passenger account,
  # trip distance, trip duration, pick up day time,payment type
class(merged1_clean$pickup_datetime)

vendor <- as.factor(merged1_clean$vendor_id)
passenger_count <- as.factor(merged1_clean$passenger_count)
head(passenger_count)
trip_dist <- as.numeric(merged1_clean$trip_distance)
trip_durat <- as.numeric(merged1_clean$trip_time_in_secs)
pickup_time <- as.factor(merged1_clean$pickup_datetime)
payment_type <- as.factor(merged1_clean$payment_type)
fare <- as.numeric(merged1_clean$fare_amount)

fare_df <- data.frame(vendor, passenger_count, trip_dist, trip_durat,
                      payment_type, fare)
head(fare_df)

data_fare <-
log(fare)~vendor+passenger_count+trip_dist+trip_durat+payment_type

model_fare <- lm(data_fare, data = fare_df)
model_fare
```

Smart Real State

```
library(readr)
library(data.table)
library(jsonlite)
library(rworldmap)
library(lubridate)
library(ggplot2)
library(prophet)
library(ggmap)

data_UK <- fread("R:/Data_Science/UK_Housing_Price.csv")
index_dataUK <- sample(1:nrow(data_UK), 30000)
prices_UK <- data_UK[index_dataUK,]

prices_UK_data<- prices_UK

save(prices_UK_data, pcodes, clean_pcodes, clean_pcodes_UK, merged_UK,
pcodes, prices_UK_data, df, m, sorted_ds, future_prices, forecast, file =
Set.Rdata")

load("set.Rdata")

head(prices_UK_data)
fix(prices_UK_data)
summary(prices_UK_data)

# Price distribution

hist(log(prices_UK_data$Price),
     xlab = "log(Price)",
     main = "Distribution of Price",
     col = "blue")

exp(0)
= 162755
Mean = 181895

# Relationship between (N=Old Y=New) and Duration(F=Freehold, L=leasehold U=
unidentified)

table(prices_UK_data$`Old/New`, prices_UK_data$Duration)

boxplot(log(prices_UK_data$Price)~prices_UK_data$Duration, xlab = "Property
Tenure",
        ylab = "log(Price)", main = "Property Tenure Vs Price",
```

```
      las = 1)

boxplot(log(prices_UK_data$Price)~prices_UK_data$`Old/New`, xlab = "Old/New",
      ylab = "log(Price)", main = "Property Age Vs Price",
      las = 1)


# Relationship between Property Type(D=Detached, S=Semi, T=Terrace, F=Flats
O= Other) and (N=Old Y=New) )

table(prices_UK_data$`Old/New`, prices_UK_data$`Property Type`)

boxplot(log(prices_UK_data$Price)~prices_UK_data$`Property Type`, xlab =
"Property Type",
      ylab = "log(Price)", main = "Property Type Vs Price",
      las = 1)

# Relationship between Property Type(D=Detached, S=Semi, T=Terrace, F=Flats
O= Other) and Duration(F=Freehold, L=leasehold U= unidentified)

table(prices_UK_data$`Property Type`, prices_UK_data$Duration)



# We want to determine how significant the price can be explained by variables
(Age, Duration, Property_t)

Age <- as.factor(prices_UK_data$`Old/New`)
is.factor(Age)
factor(Age)

Duration <- as.factor(prices_UK_data$Duration)
is.factor(Duration)
factor(Duration)

Property_t <- as.factor(prices_UK_data$`Property Type`)
is.factor(Property_t)
factor(Property_t)

District <- as.factor(prices_UK_data$District)
is.factor(District)
factor(District)

Price <- prices_UK_data$Price
class(Price)

factor_data <- data.frame(Price, Age, Duration, Property_t)
```

```
head(factor_data)


dp <- log(Price) ~ Age + Duration + Property_t

model1 <- lm(dp, data=factor_data)
summary(model1)


# creating a data frame(df) with only two columns; Date of Transfer(ds) and
Price(y)

ds <- as.Date(prices_UK_data$`Date of Transfer`, "%Y-%m-%d")

y <- log(prices_UK_data$Price)

df <- data.frame(ds,y)
head(df)

# sorting by date of transfer

sorted_ds <- df[order(ds),]

head(sorted_ds)
summary(sorted_ds)

qplot(ds,y)

View(sorted_ds)

# Incorporating prophet package developed by Facebook for prediction of 365 days
m<-prophet(sorted_ds)

future_prices <- make_future_dataframe(m, periods = 365)

tail(future_prices)

forecast <- predict(m, future_prices)

tail(forecast)

tail(forecast[c("ds", "yhat", "yhat_lower", "yhat_upper")])


# Example. Prediction for 2018-06-23

exp(12.36796) = 235146
```

```
##############################################
# Warning for plotting, it can crash the software
plot(m, forecast)
########################################

# Componenets of timeseries
prophet_plot_components(m,forecast)



# mapping for most expensive (>75% quantile) and cheapest properties (< 25%
quantile)

summary(prices_UK_data)
str(prices_UK_data)

pcodes <- fromJSON("R:/Data_Science/postcodes.json")

clean_pcodes <- pcodes[!duplicated(pcodes$town),]

town_and_regions <- paste(pcodes$town,pcodes$region)

clean_pcodes_UK <- pcodes[!duplicated(town_and_regions),]

clean_pcodes_UK$town <- toupper(clean_pcodes_UK$town)

clean_pcodes$town <- toupper(clean_pcodes$town)

merged_UK <- merge(x = prices_UK_data, y = clean_pcodes_UK,
                   by.x=c("Town/City"), by.y=c("town"), all = FALSE)

str(merged_UK)
summary(merged_UK)
fix(merged_UK)

quantile(prices_UK_data$Price, probs = c(0, 0.25, 0.50, 0.75, 1))

expensive_merged_UK <- merged_UK[merged_UK[,3] >= 210000,]

View(expensive_merged_UK)
str(expensive_merged_UK)
summary(expensive_merged_UK)

cheap_merged_UK <- merged_UK[merged_UK[,3]<= 75000,]
str(cheap_merged_UK)

# Selecting the columns we need. Price, longitude, latitude
```

```
expensive_merged_UK <- expensive_merged_UK[,c(3,16,19)]

cheap_merged_UK <- cheap_merged_UK[,c(3,16,19)]

map_UK <- get_map("England", zoom =7, maptype = "hybrid", source = "google")


# converting longitud and latitude columns to a numeric value
expensive_numeric_longitude <- as.numeric(expensive_merged_UK$longitude)

expensive_numeric_latitude <- as.numeric(expensive_merged_UK$latitude)

cheap_numeric_longitude <- as.numeric(cheap_merged_UK$longitude)

cheap_numeric_latitude <- as.numeric(cheap_merged_UK$latitude)


ggmap(map_UK) + geom_point(data=expensive_merged_UK, aes(x =
expensive_numeric_longitude, y = expensive_numeric_latitude), color = "green",
size = 2, alpha = 0.2)+
  geom_point(data=cheap_merged_UK, aes(x = cheap_numeric_longitude, y =
cheap_numeric_latitude), color = "blue", size = 2, alpha = 0.2)
```