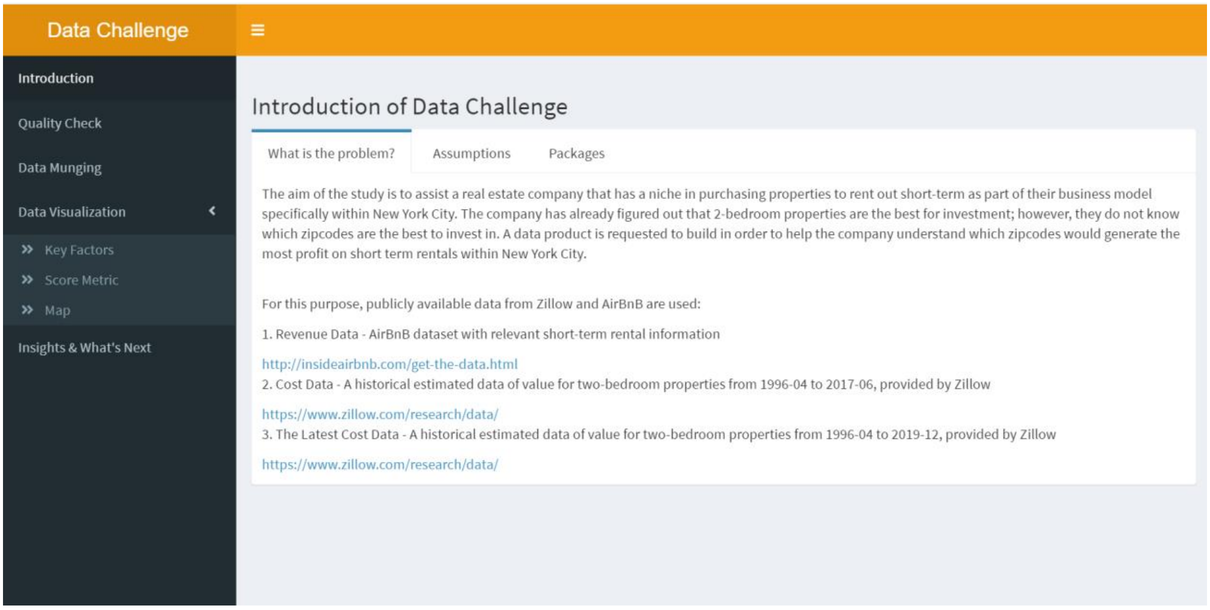
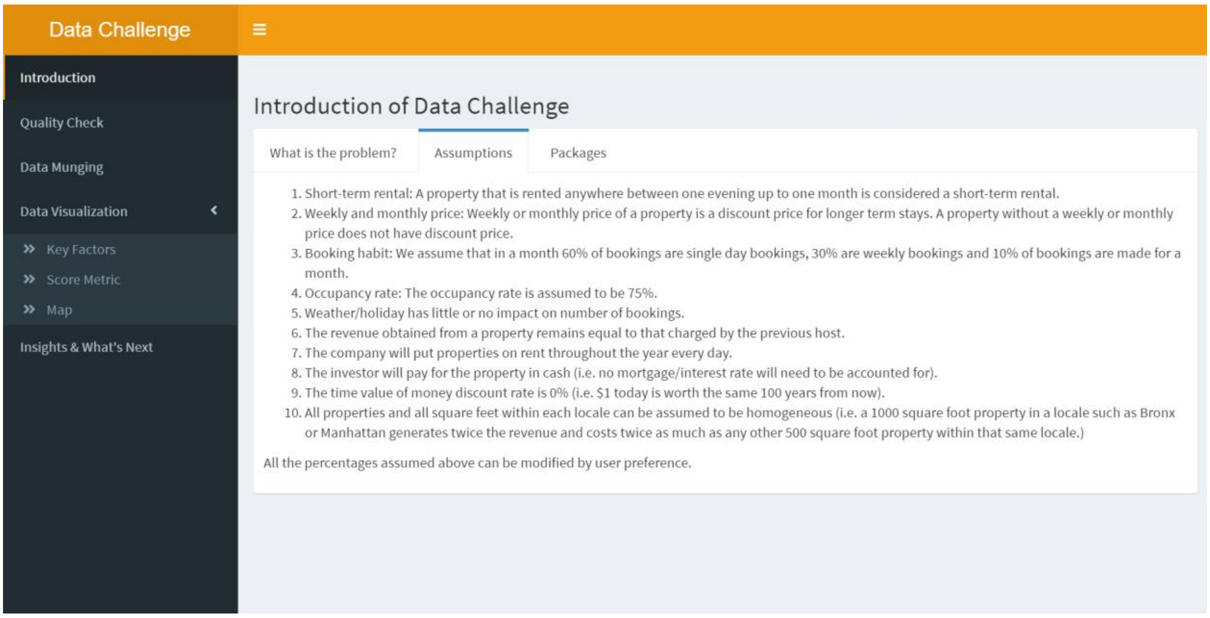
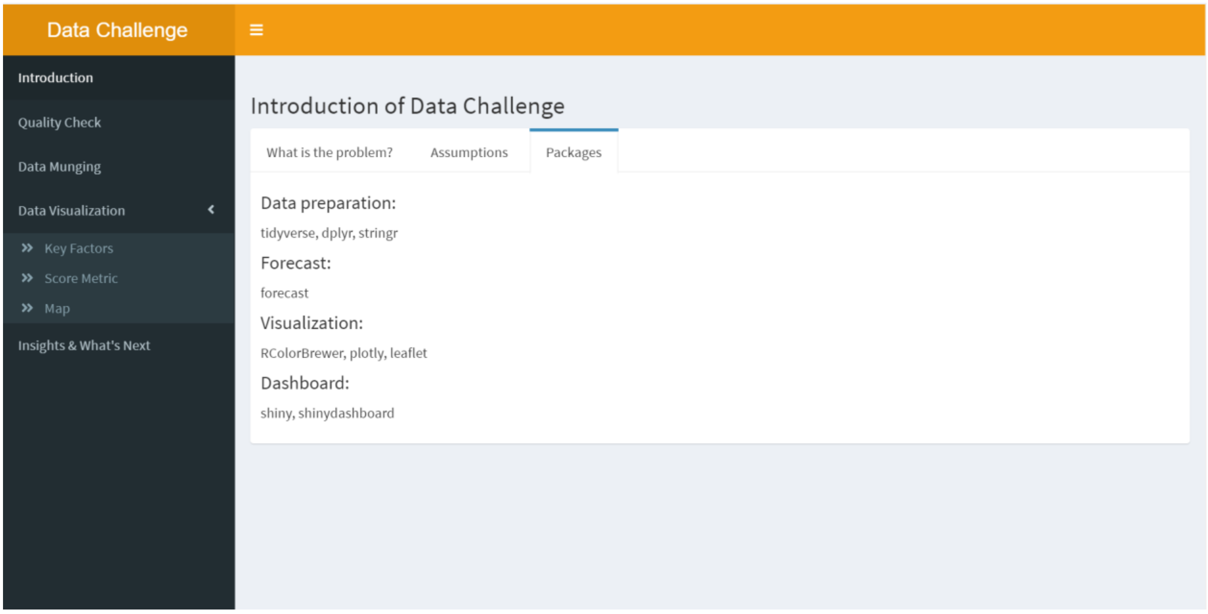
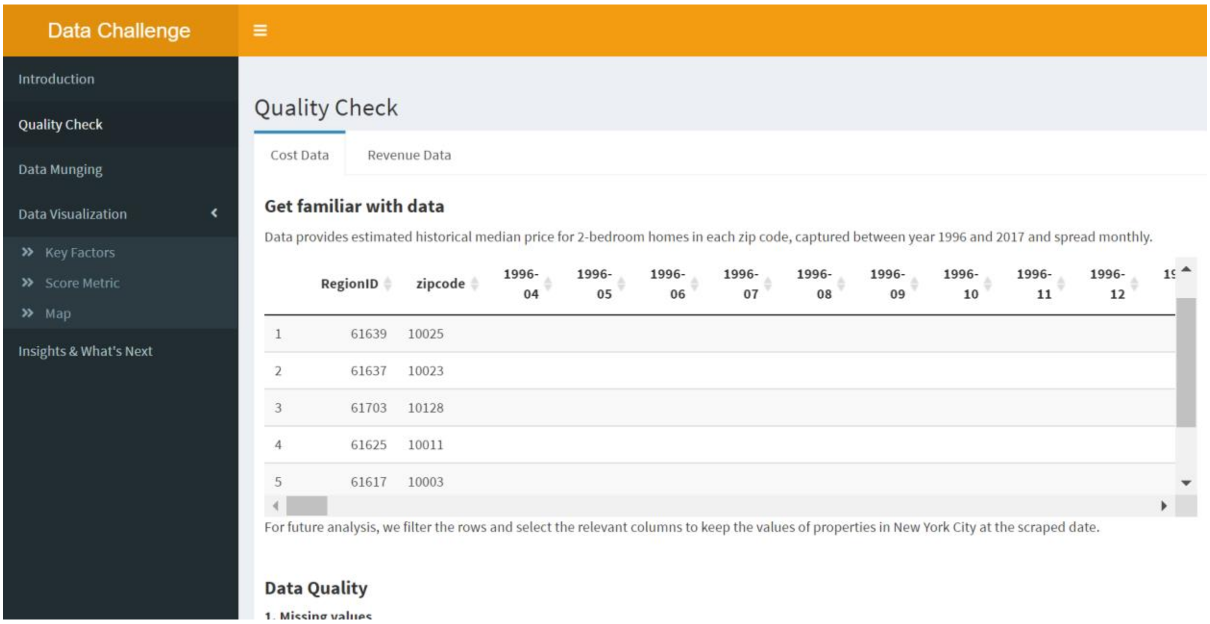

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Introduction of Data Challenge
What is the problem?
Assumptions
Packages
The aim of the study is to assist a real estate company that has a niche in purchasing properties to rent out short-term as part of their business model specifically within New York City. The company has already figured out that 2-bedroom properties are the best for investment; however, they do not know which zipcodes are the best to invest in. A data product is requested to build in order to help the company understand which zipcodes would generate the most profit on short term rentals within New York City.
For this purpose, publicly available data from Zillow and AirBnB are used:
1. Revenue Data - AirBnB dataset with relevant short-term rental information
http://insideairbnb.com/get-the-data.html
2. Cost Data - A historical estimated data of value for two-bedroom properties from 1996-04 to 2017-06, provided by Zillow
https://www.zillow.com/research/data/
3. The Latest Cost Data - A historical estimated data of value for two-bedroom properties from 1996-04 to 2019-12, provided by Zillow
https://www.zillow.com/research/data/


Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Introduction of Data Challenge
What is the problem?
Assumptions
Packages
1. Short-term rental: A property that is rented anywhere between one evening up to one month is considered a short-term rental.
2. Weekly and monthly price: Weekly or monthly price of a property is a discount price for longer term stays. A property without a weekly or monthly price does not have discount price.
3. Booking habit: We assume that in a month 60% of bookings are single day bookings, 30% are weekly bookings and 10% of bookings are made for a month.
4. Occupancy rate: The occupancy rate is assumed to be 75%.
5. Weather/holiday has little or no impact on number of bookings.
6. The revenue obtained from a property remains equal to that charged by the previous host.
7. The company will put properties on rent throughout the year every day.
8. The investor will pay for the property in cash (i.e. no mortgage/interest rate will need to be accounted for).
9. The time value of money discount rate is 0% (i.e. $1 today is worth the same 100 years from now).
10. All properties and all square feet within each locale can be assumed to be homogeneous (i.e. a 1000 square foot property in a locale such as Bronx or Manhattan generates twice the revenue and costs twice as much as any other 500 square foot property within that same locale.)
All the percentages assumed above can be modified by user preference.


Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Introduction of Data Challenge
What is the problem?
Assumptions
Packages
Data preparation:
tidyverse, dplyr, stringr
Forecast:
forecast
Visualization:
RColorBrewer, plotly, leaflet
Dashboard:
shiny, shinydashboard


Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Quality Check
Cost Data
Revenue Data
Get familiar with data
Data provides estimated historical median price for 2-bedroom homes in each zip code, captured between year 1996 and 2017 and spread monthly.
RegionID
zipcode
1996-04
1996-05
1996-06
1996-07
1996-08
1996-09
1996-10
1996-11
1996-12
1 61639 10025
2 61637 10023
3 61703 10128
4 61625 10011
5 61617 10003
For future analysis, we filter the rows and select the relevant columns to keep the values of properties in New York City at the scraped date.
Data Quality
1. Missing values

Showing 1 to 6 of 6 entries Previous 1 Next

For future analysis, we filter the rows and select the relevant columns to keep the values of properties in New York City at the scraped date.

Data Quality

1. Missing values

Median price for early years (1996-2013) has plenty of NAs as shown in the table below. Steps are taken in the following section to exclude columns with NAs when trend calculation and time series forecast.

Show 10 entries Search:

	cost_missing_val
RegionID	0
zipcode	0
1996-04	17
1996-05	17

2. Missing zipcodes of NYC

New York hosts 176 zipcodes. However, there are only 25 zipcodes of NYC recorded in this data set. We are going to settle with this data for now. New data source can be connected in the future to account for rest of the zipcodes.

Data Challenge

- Introduction
- Quality Check
- Data Munging
- Data Visualization
  - Key Factors
  - Score Metric
  - Map
- Insights & What's Next

Quality Check

Cost Data | Revenue Data

Get familiar with data

Revenue data contains information including details about the properties including location, number of bedrooms, room types, price and other details for stay.

Show 10 entries Search:

id	listing_url	scrape_id	last_scraped	name	summary	space

In order to simplify the analysis, we only select 15 relevant variables and filter 2-bedroom properties in NYC to create a subset for the following section.

Data Quality

1. Needless character

Data Quality

1. Needless character

'$' Value prefix of every price row prevents numeric manipulation. It is thus removed from three columns: Price, Weekly Price & Monthly Price.

2. Missing values

Show 10 entries Search:

	rev_missing_val
id	0
listing_url	0
scrape_id	0
last_scraped	0

(1) zipcode

There are less than 1% missing values in zipcode. Ignoring these values will not cause much effect to the analysis.

(2) weekly and monthly price

There are more than 80% weekly and monthly price missing. Since we assume that the weekly or monthly price of a property is a discount price for longer term stays, and a property without a weekly or monthly price does not have discount price, we can simply fill in the NAs with 7 and 31 times of daily price.

3. Extreme values

There are a few extreme values in price such as $9,000 or $0 per night. The reason for this issue could be a data input error of the hosts. We can simply remove these rows.

4. Ambiguous match between zipcode and neighborhood

4. Ambiguous match between zipcode and neighborhood

Some zipcodes belong to more than one neighborhood in the dataset. To solve this problem, we keep the neighborhood to which the zipcode usually belongs as the only match.

Data Challenge

Introduction

Quality Check

Data Munging

Data Visualization

Key Factors

Score Metric

Map

Insights & What's Next

Data Munging

Forecast | Data Join | Price Corrected

Time Unit

The Zillow and AirBnb datasets have different units of time. In order to operate the analysis, we determine a common unit of time - month.

ARIMA Forecast

The revenue data is scraped in 2019-07, but the cost data only have historical house value from 1996-04 to 2017-06. Due to shortage of time, we assume that there is seasonality in the price and that values depend not only on previous values (Auto Regressive AR) but also on differences between previous values (Moving Average MA). So, we apply, **ARIMA** model to predict the cost of the properties in zipcodes from 2017-07 to 2019-07. We then attach the price of property in 2019-07 calculated at zipcode level with each zipcode as a new column to cost data.

Data Challenge

Introduction

Quality Check

Data Munging

Data Visualization

Key Factors

Score Metric

Map

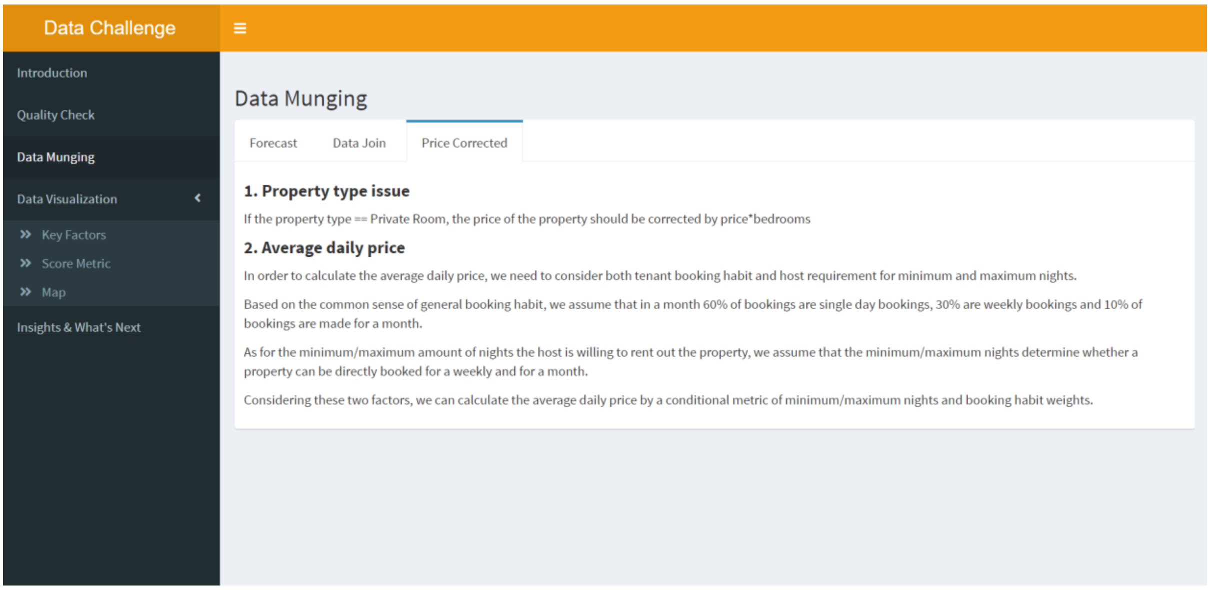
Insights & What's Next

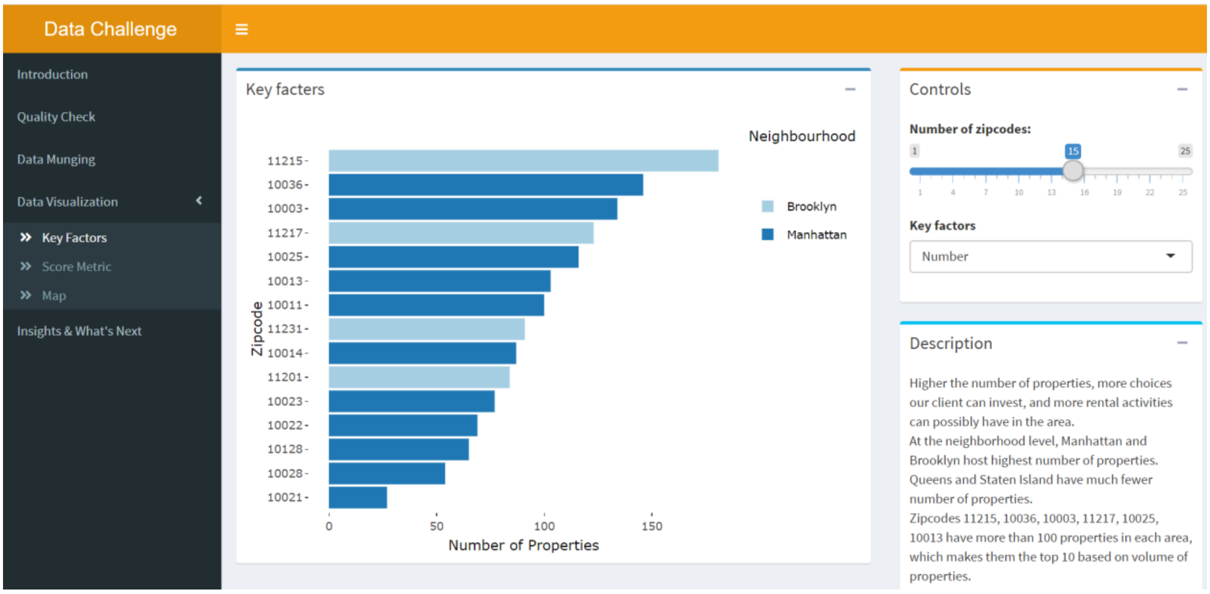
Data Munging

Forecast | Data Join | Price Corrected

In order to make sure this product is always applicable whenever new data is available or whenever we are ready to approach a new market, a function is built to link the data together in a scalable way by matching the scraped date and zipcode of two datasets.

After inner join, we recognize that there are 1508 properties of 24 zipcodes in 4 distinct neighborhoods - Manhattan, Queens, Brooklyn and Staten Island - matching together.

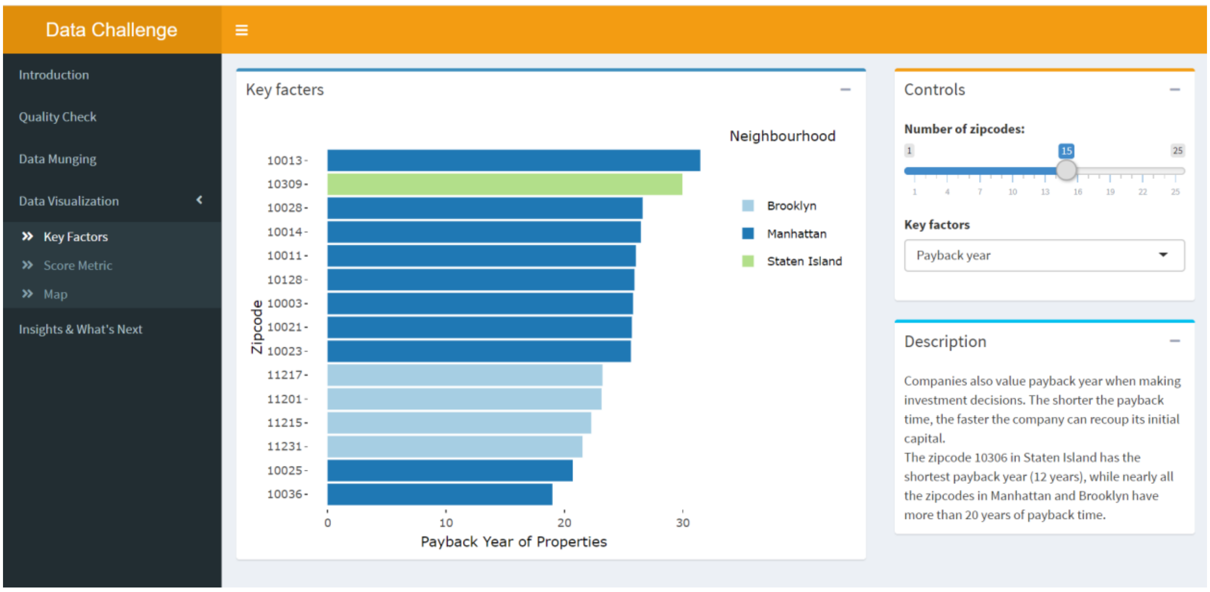

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Data Munging
Forecast
Data Join
Price Corrected
1. Property type issue
If the property type == Private Room, the price of the property should be corrected by price*bedrooms
2. Average daily price
In order to calculate the average daily price, we need to consider both tenant booking habit and host requirement for minimum and maximum nights.
Based on the common sense of general booking habit, we assume that in a month 60% of bookings are single day bookings, 30% are weekly bookings and 10% of bookings are made for a month.
As for the minimum/maximum amount of nights the host is willing to rent out the property, we assume that the minimum/maximum nights determine whether a property can be directly booked for a weekly and for a month.
Considering these two factors, we can calculate the average daily price by a conditional metric of minimum/maximum nights and booking habit weights.

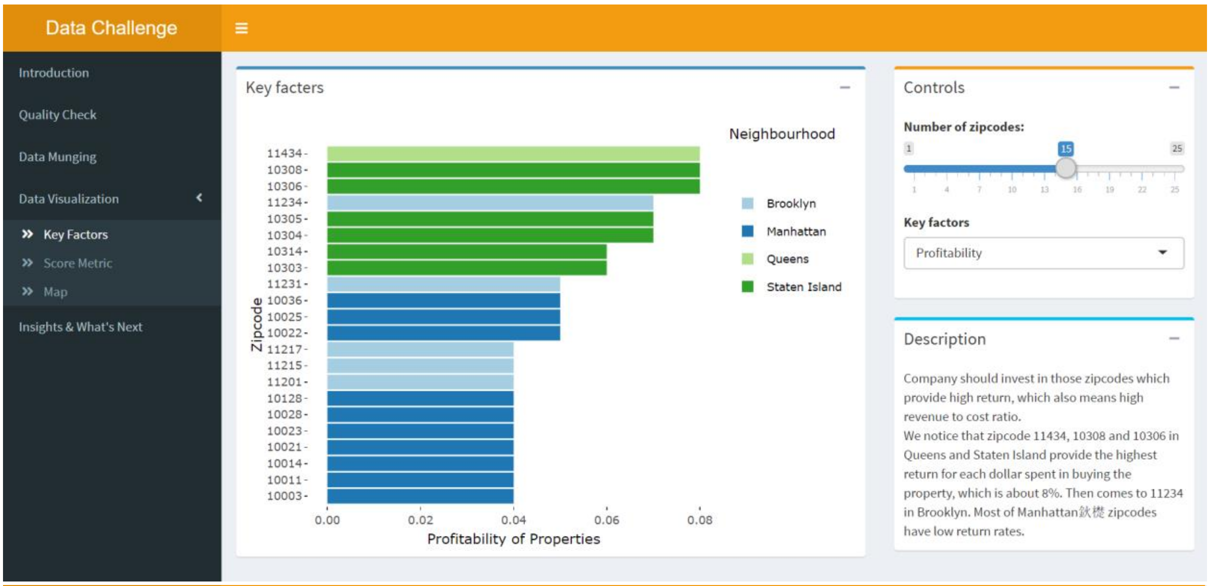

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Key facters
Neighbourhood
Brooklyn
Manhattan
11215
10036
10003
11217
10025
10013
10011
11231
10014
11201
10023
10022
10128
10028
10021
Zipcode
0
50
100
150
Number of Properties
Controls
Number of zipcodes:
1
15
25
Key factors
Number
Description
Higher the number of properties, more choices our client can invest, and more rental activities can possibly have in the area.
At the neighborhood level, Manhattan and Brooklyn host highest number of properties. Queens and Staten Island have much fewer number of properties.
Zipcodes 11215, 10036, 10003, 11217, 10025, 10013 have more than 100 properties in each area, which makes them the top 10 based on volume of properties.

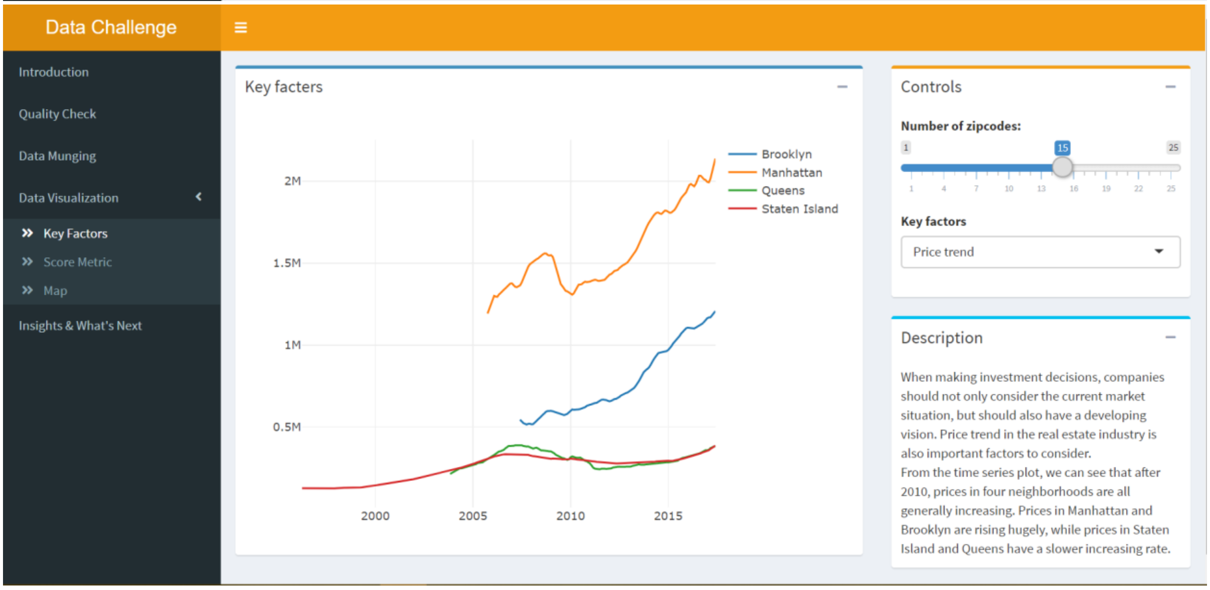

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Key facters
Neighbourhood
Brooklyn
Manhattan
10013
10011
10014
10003
10028
10023
10021
10128
10022
10036
10025
11201
11217
11231
11215
Zipcode
0
1M
2M
Cost of Properties
Controls
Number of zipcodes:
1
15
25
Key factors
Cost
Description
One of the primary constraints in decision of investment is the cost. Lower the cost, company can save a lot more and reach ROI faster.
Zipcodes in Manhattan walk away with highest property cost, with an average of about $2M.
Brooklyn comes second, with about $1M.
Zipcodes in Staten Island and Queens have much lower prices, with lower than $0.5M.

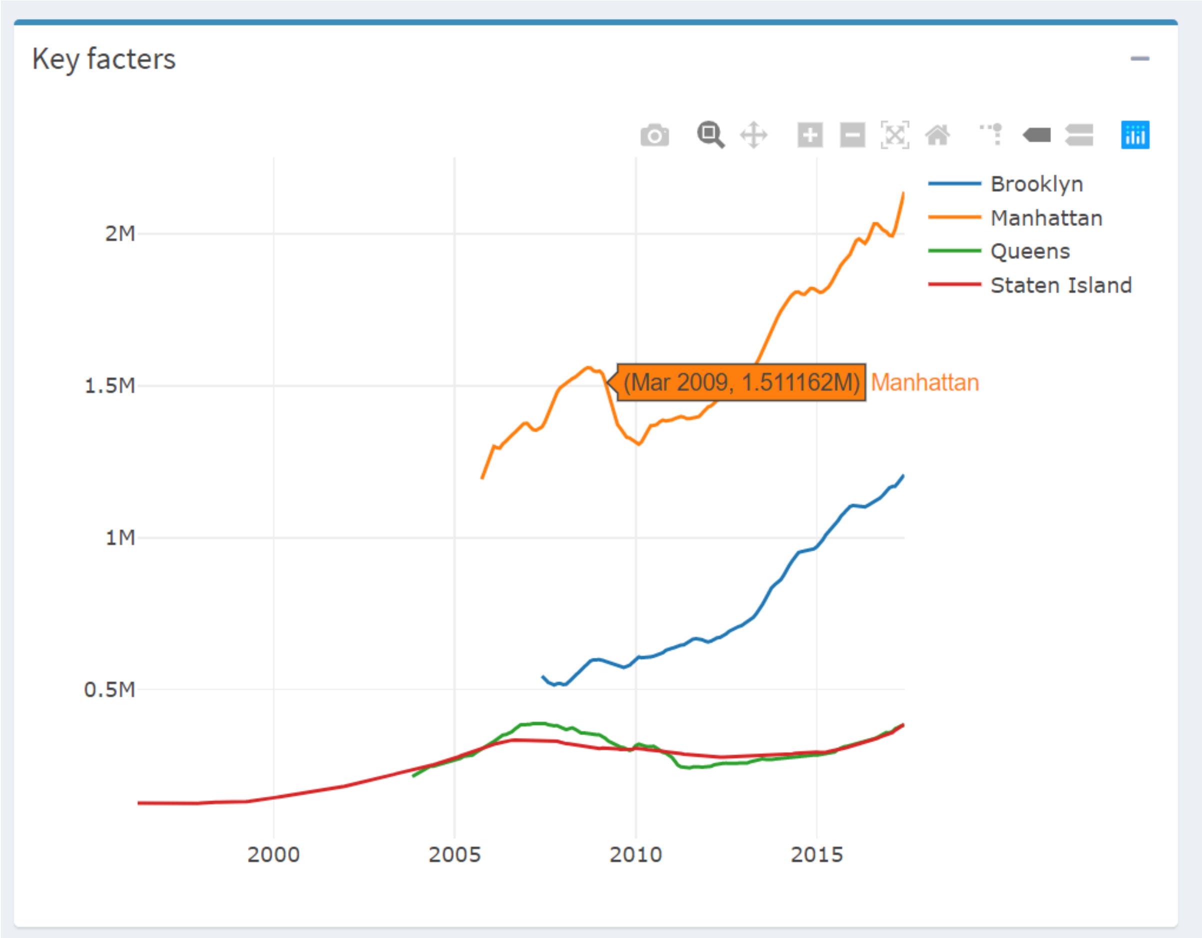

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Key facters
Neighbourhood
Brooklyn
Manhattan
Queens
Staten Island
Zipcode
Daily Price of Properties
Controls
Number of zipcodes:
Key factors
Nightly price
Description
Higher the average daily rental price, faster is the cash flow. Therefore, company would look forward to making quick bucks with high daily price.
Manhattan has a wider spread with higher price ranging to $50-$1000 per night. Brooklyn comes second also with wide spread but lower price comparing to Manhattan. Staten Island and Queens have lower price and narrower distribution, and this is probably because of limited sample size.

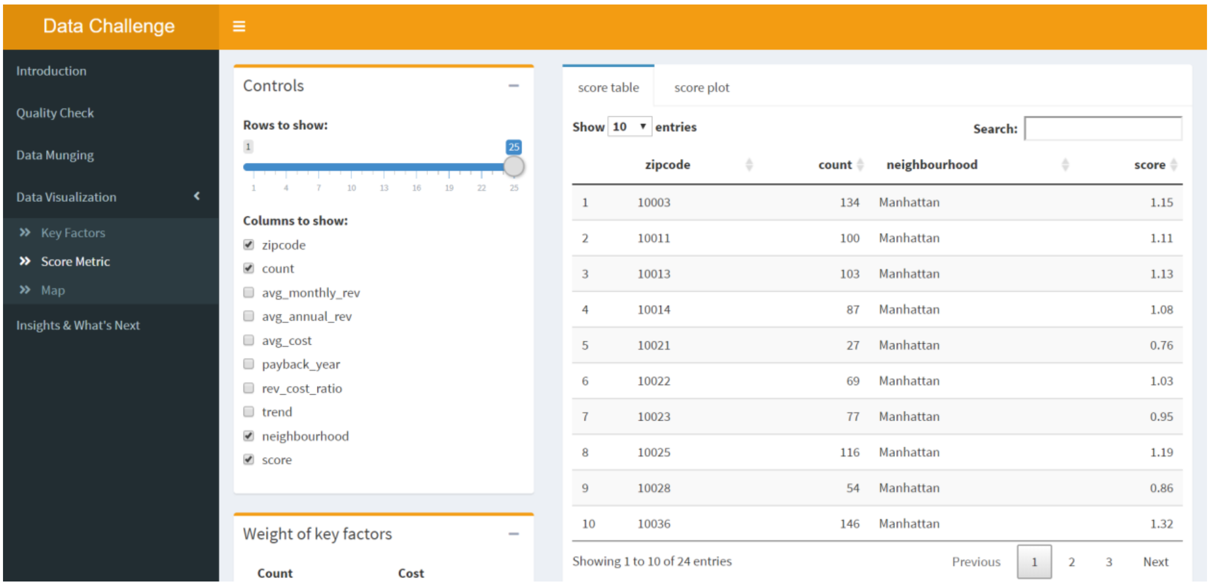

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Key facters
Neighbourhood
Brooklyn
Manhattan
Zipcode
Annual Revenue of Properties
Controls
Number of zipcodes:
Key factors
Annual revenue
Description
Different from the daily price, the annual revenue also taken into account the effects of weekly/monthly discount prices, minimum/maximum days and occupancy rate. Therefore, the annual revenue will provide a more comprehensive information to help us make better investment choices.
As usual, Manhattan has the highest revenue, and Brooklyn comes second.

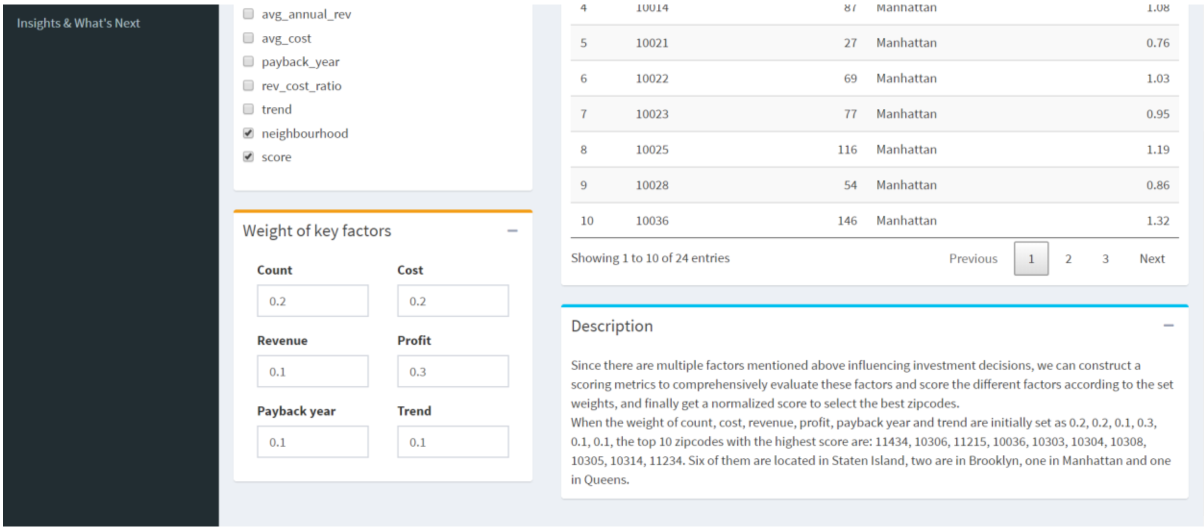

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Key facters
Neighbourhood
Brooklyn
Manhattan
Staten Island
Zipcode
Payback Year of Properties
Controls
Number of zipcodes:
Key factors
Payback year
Description
Companies also value payback year when making investment decisions. The shorter the payback time, the faster the company can recoup its initial capital.
The zipcode 10306 in Staten Island has the shortest payback year (12 years), while nearly all the zipcodes in Manhattan and Brooklyn have more than 20 years of payback time.

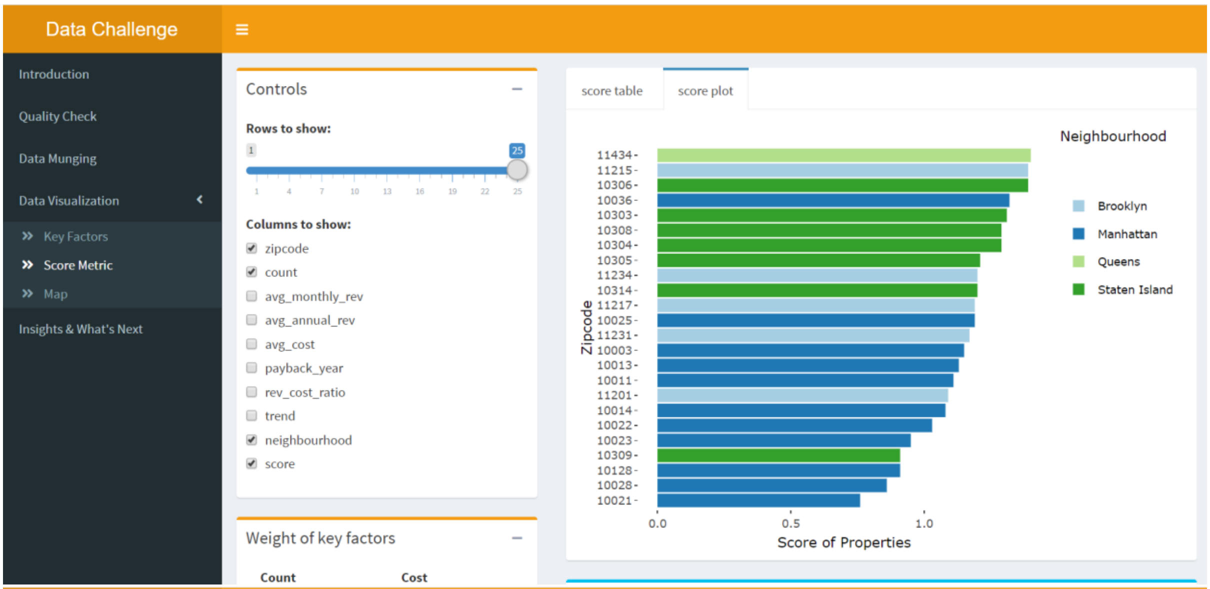

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Key facters
Neighbourhood
Brooklyn
Manhattan
Queens
Staten Island
Zipcode
11434
10308
10306
11234
10305
10304
10314
10303
11231
10036
10025
10022
11217
11215
11201
10128
10028
10023
10021
10014
10011
10003
0.00
0.02
0.04
0.06
0.08
Profitability of Properties
Controls
Number of zipcodes:
Key factors
Profitability
Description
Company should invest in those zipcodes which provide high return, which also means high revenue to cost ratio.
We notice that zipcode 11434, 10308 and 10306 in Queens and Staten Island provide the highest return for each dollar spent in buying the property, which is about 8%. Then comes to 11234 in Brooklyn. Most of Manhattan鈥檚 zipcodes have low return rates.

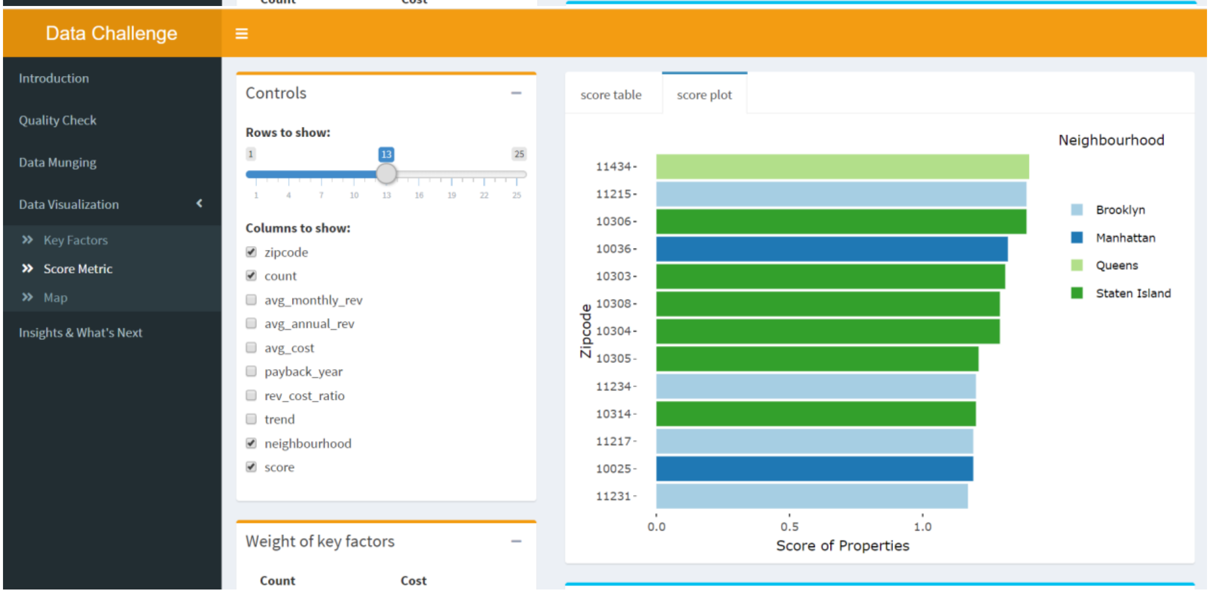

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Key facters
Brooklyn
Manhattan
Queens
Staten Island
2M
1.5M
1M
0.5M
2000
2005
2010
2015
Controls
Number of zipcodes:
Key factors
Price trend
Description
When making investment decisions, companies should not only consider the current market situation, but should also have a developing vision. Price trend in the real estate industry is also important factors to consider.
From the time series plot, we can see that after 2010, prices in four neighborhoods are all generally increasing. Prices in Manhattan and Brooklyn are rising hugely, while prices in Staten Island and Queens have a slower increasing rate.

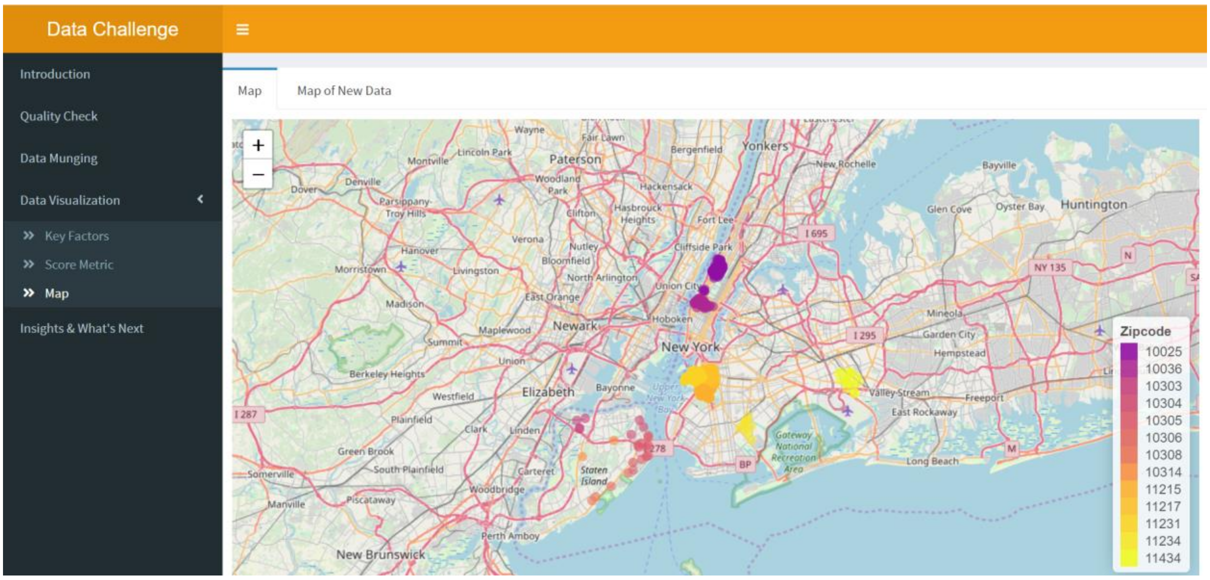

Key facters
2M
1.5M
1M
0.5M
2000
2005
2010
2015
Brooklyn
Manhattan
Queens
Staten Island
(Mar 2009, 1.511162M) Manhattan

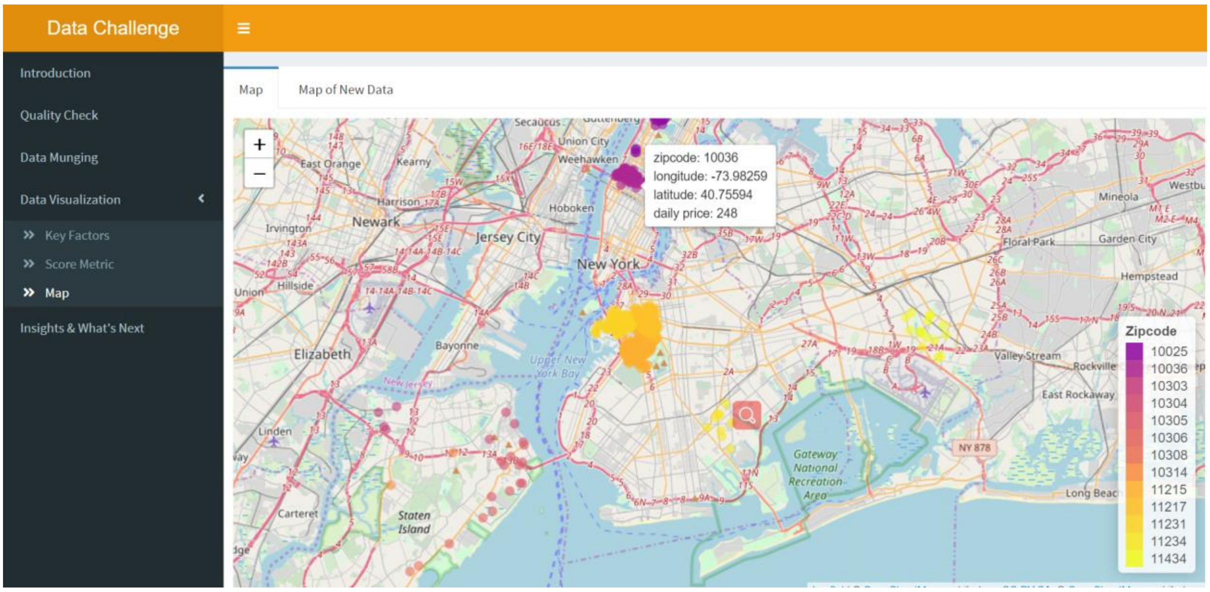

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Controls
Rows to show:
Columns to show:
zipcode
count
avg_monthly_rev
avg_annual_rev
avg_cost
payback_year
rev_cost_ratio
trend
neighbourhood
score
Weight of key factors
Count
Cost
score table
score plot
Show 10 entries
Search:
zipcode count neighbourhood score
1 10003 134 Manhattan 1.15
2 10011 100 Manhattan 1.11
3 10013 103 Manhattan 1.13
4 10014 87 Manhattan 1.08
5 10021 27 Manhattan 0.76
6 10022 69 Manhattan 1.03
7 10023 77 Manhattan 0.95
8 10025 116 Manhattan 1.19
9 10028 54 Manhattan 0.86
10 10036 146 Manhattan 1.32
Showing 1 to 10 of 24 entries
Previous
1
2
3
Next

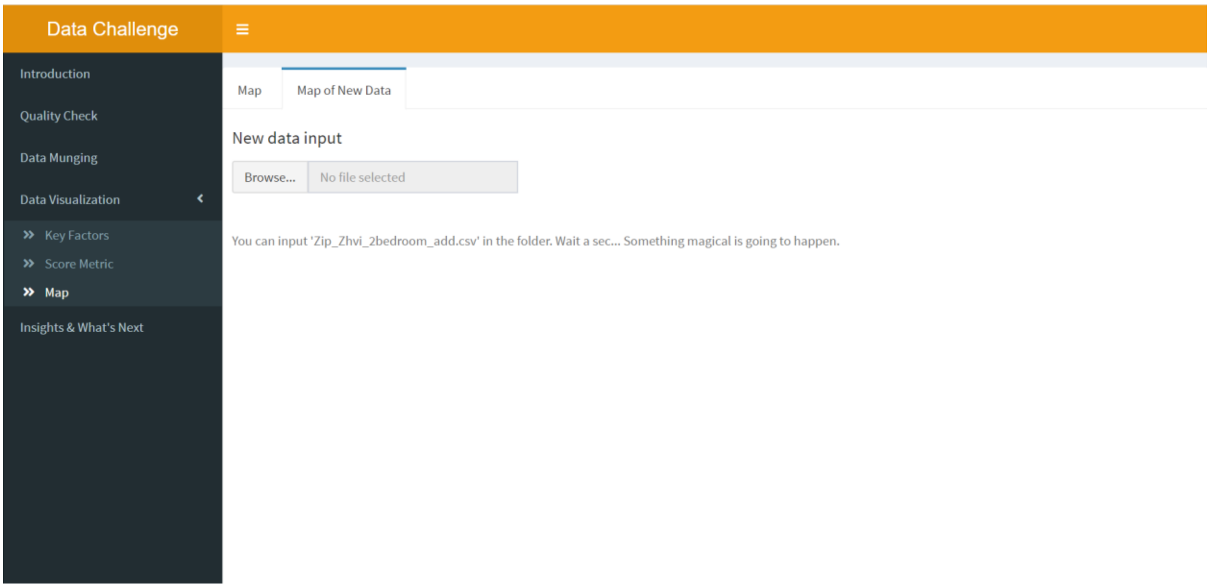

Insights & What's Next
avg_annual_rev
avg_cost
payback_year
rev_cost_ratio
trend
neighbourhood
score
Weight of key factors
Count
0.2
Cost
0.2
Revenue
0.1
Profit
0.3
Payback year
0.1
Trend
0.1
4 10014 87 Manhattan 1.08
5 10021 27 Manhattan 0.76
6 10022 69 Manhattan 1.03
7 10023 77 Manhattan 0.95
8 10025 116 Manhattan 1.19
9 10028 54 Manhattan 0.86
10 10036 146 Manhattan 1.32
Showing 1 to 10 of 24 entries
Previous 1 2 3 Next
Description
Since there are multiple factors mentioned above influencing investment decisions, we can construct a scoring metrics to comprehensively evaluate these factors and score the different factors according to the set weights, and finally get a normalized score to select the best zipcodes.
When the weight of count, cost, revenue, profit, payback year and trend are initially set as 0.2, 0.2, 0.1, 0.3, 0.1, 0.1, the top 10 zipcodes with the highest score are: 11434, 10306, 11215, 10036, 10303, 10304, 10308, 10305, 10314, 11234. Six of them are located in Staten Island, two are in Brooklyn, one in Manhattan and one in Queens.

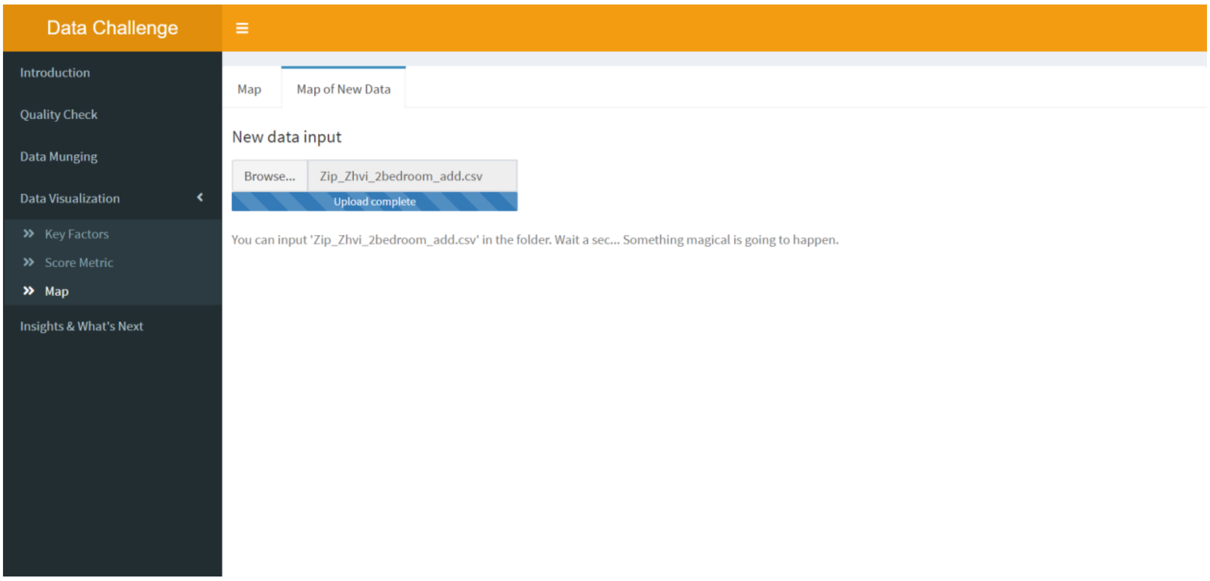

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Controls
Rows to show:
1
25
Columns to show:
zipcode
count
avg_monthly_rev
avg_annual_rev
avg_cost
payback_year
rev_cost_ratio
trend
neighbourhood
score
Weight of key factors
Count
Cost
score table
score plot
Neighbourhood
Brooklyn
Manhattan
Queens
Staten Island
Zipcode
11434
11215
10306
10036
10303
10308
10304
10305
11234
10314
11217
10025
11231
10003
10013
10011
11201
10014
10022
10023
10309
10128
10028
10021
0.0
0.5
1.0
Score of Properties

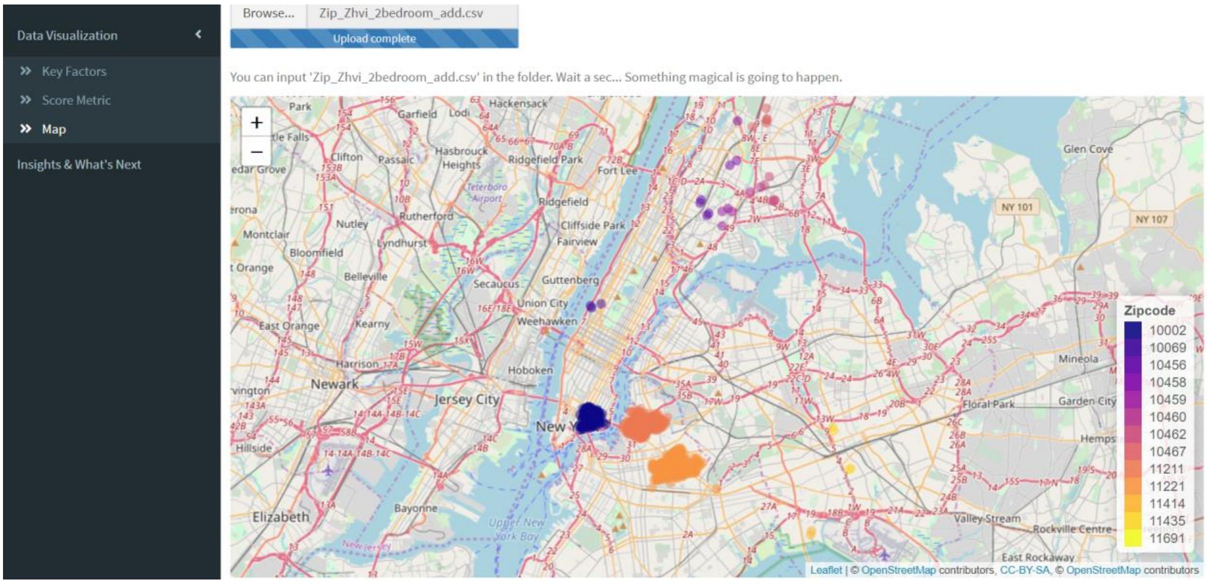

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Controls
Rows to show:
1
13
25
Columns to show:
zipcode
count
avg_monthly_rev
avg_annual_rev
avg_cost
payback_year
rev_cost_ratio
trend
neighbourhood
score
Weight of key factors
Count
Cost
score table
score plot
Neighbourhood
Brooklyn
Manhattan
Queens
Staten Island
Zipcode
11434
11215
10306
10036
10303
10308
10304
10305
11234
10314
11217
10025
11231
0.0
0.5
1.0
Score of Properties

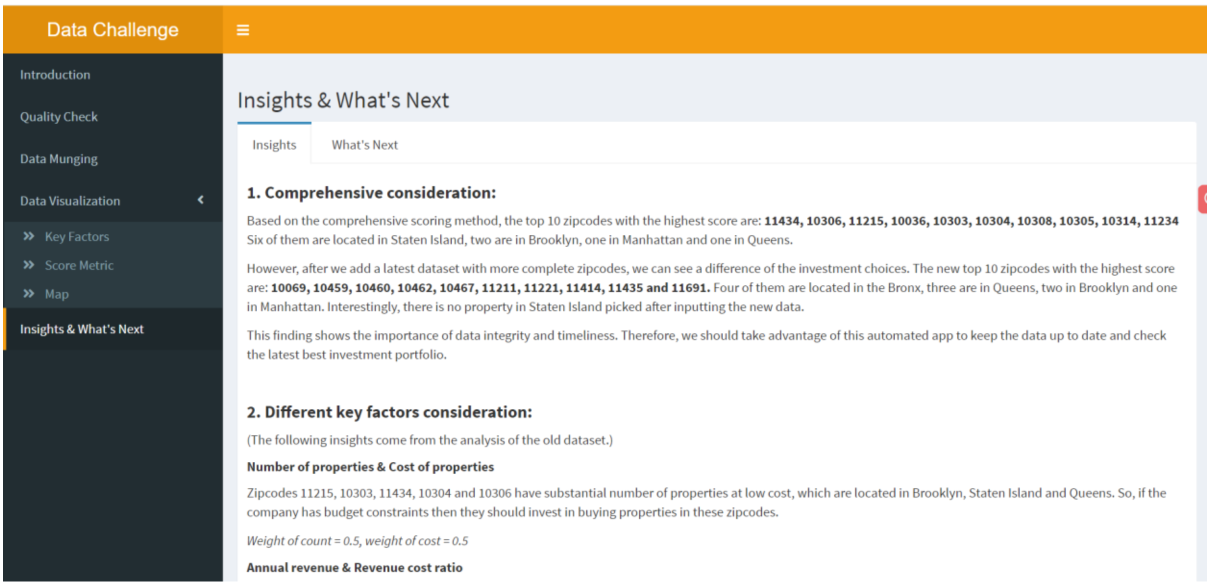

Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Map
Map of New Data
Zipcode
10025
10036
10303
10304
10305
10306
10308
10314
11215
11217
11231
11234
11434


Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Map
Map of New Data
zipcode: 10036
longitude: -73.98259
latitude: 40.75594
daily price: 248
Zipcode
10025
10036
10303
10304
10305
10306
10308
10314
11215
11217
11231
11234
11434


Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Map
Map of New Data
New data input
Browse...
No file selected
You can input 'Zip_Zhvi_2bedroom_add.csv' in the folder. Wait a sec... Something magical is going to happen.


Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Map
Map of New Data
New data input
Browse...
Zip_Zhvi_2bedroom_add.csv
Upload complete
You can input 'Zip_Zhvi_2bedroom_add.csv' in the folder. Wait a sec... Something magical is going to happen.


Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Browse...
Zip_Zhvi_2bedroom_add.csv
Upload complete
You can input 'Zip_Zhvi_2bedroom_add.csv' in the folder. Wait a sec... Something magical is going to happen.
Zipcode
10002
10069
10456
10458
10459
10460
10462
10467
11211
11221
11414
11435
11691
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap contributors


Data Challenge
Introduction
Quality Check
Data Munging
Data Visualization
Key Factors
Score Metric
Map
Insights & What's Next
Insights & What's Next
Insights
What's Next
1. Comprehensive consideration:
Based on the comprehensive scoring method, the top 10 zipcodes with the highest score are: 11434, 10306, 11215, 10036, 10303, 10304, 10308, 10305, 10314, 11234
Six of them are located in Staten Island, two are in Brooklyn, one in Manhattan and one in Queens.
However, after we add a latest dataset with more complete zipcodes, we can see a difference of the investment choices. The new top 10 zipcodes with the highest score are: 10069, 10459, 10460, 10462, 10467, 11211, 11221, 11414, 11435 and 11691. Four of them are located in the Bronx, three are in Queens, two in Brooklyn and one in Manhattan. Interestingly, there is no property in Staten Island picked after inputting the new data.
This finding shows the importance of data integrity and timeliness. Therefore, we should take advantage of this automated app to keep the data up to date and check the latest best investment portfolio.
2. Different key factors consideration:
(The following insights come from the analysis of the old dataset.)
Number of properties & Cost of properties
Zipcodes 11215, 10303, 11434, 10304 and 10306 have substantial number of properties at low cost, which are located in Brooklyn, Staten Island and Queens. So, if the company has budget constraints then they should invest in buying properties in these zipcodes.
Weight of count = 0.5, weight of cost = 0.5
Annual revenue & Revenue cost ratio

2. Different key factors consideration:

(The following insights come from the analysis of the old dataset.)

Number of properties & Cost of properties

Zipcodes 11215, 10303, 11434, 10304 and 10306 have substantial number of properties at low cost, which are located in Brooklyn, Staten Island and Queens. So, if the company has budget constraints then they should invest in buying properties in these zipcodes.

Weight of count = 0.5, weight of cost = 0.5

Annual revenue & Revenue cost ratio

If the company is willing to buy properties having high cost, then they should invest in Zipcodes 10038 in Staten Island, 11215 in Brooklyn, 10036, 10003 and 10025 in Manhattan because these zipcodes not only have high number of costly properties but also provide very high return as the revenue is high.

Weight of count = 0.5, weight of revenue = 0.2, weight of profit = 0.5

Payback year

If the company is more focused on getting its money back quickly with a limited initial investment, zipcodes 10038, 10314, 11434, 10303 and 10306 fit the bill, which are all located in Staten Island and Queens.

Weight of cost = 0.5, weight of payback year = 0.5

Price trend

If the company attaches importance on long-term growth and future earnings and wants to invest in properties where rents are rising faster, these are the zipcodes to go: 10013, 10014, 10011 in Manhattan, 10308 in Staten Island, and 11201 in Brooklyn.

Weight of profit = 0.5, weight of trend = 0.5

3. Other thoughts:

- Our basic recommendation for the company is to diversify and buy properties in top performing zips of different neighborhoods with prime focus on Staten Island.
- Through our products, the company can adjust the weight of each factor according to its actual needs to obtain the most suitable investment portfolio plan for its own investment strategy.

3. Other thoughts:

- Our basic recommendation for the company is to diversify and buy properties in top performing zips of different neighborhoods with prime focus on Staten Island.
- Through our products, the company can adjust the weight of each factor according to its actual needs to obtain the most suitable investment portfolio plan for its own investment strategy.
- If the company is a risk taker, it can choose the property with high investment and high yield. If the company is risk averse, it can choose the property with low initial cost and fast return.
- Company should diversify its investment portfolios and locations to not only minimize investment risk, but also lay the foundation for future expansion in multiple markets.

Data Challenge ≡

Introduction

Quality Check

Data Munging

Data Visualization ‹

» Key Factors

» Score Metric

» Map

Insights & What's Next

Insights & What's Next

Insights | What's Next

- New York hosts 176 zipcodes, data can be further enriched to account for rest of the zipcodes. This would give the company more opportunities to diverse the investment portfolio.
- Future revenue prediction: In the present analysis, we used average daily price * 360 * occupancy rate to calculate the annual revenue. It is true based on the assumption that the rental revenue will not dramatically change within a year. However, since the house price is continuingly increasing monthly, the rental price may also increase by time. Future analysis can consider the price trend of rental price in order to calculate income more accurately and realistically.
- Occupancy rate: In the present analysis, we simply assume the occupancy rate is 0.75 for every property in every zipcode. However, in reality, there may be many factors which can impact occupancy rate, such as review score, location and room quality. In the future, we could customize occupancy rate of each property by building a regression model.
- Interest rate: In this case, we have taken 0% discount rate as our assumption, but this assumption is not realistic. Some reasonable percentage rate can be taken to calculate NPV value and make a more accurate prediction.
- Text analytics on ignored description columns from revenue data: This would open insights about other metrics that drive customer to book an AirBnb property for rental such as access to public transport, parking space, etc.