Applied Data Science with R Capstone project

<Chen Yingxuan>

<2021.11.21>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- I am a Data Scientist who has recently joined an AI-powered weather data analytic company and be presented with a challenge that requires data collection, analysis, basic hypothesis testing, visualization, modelling, and dashboard to be performed on real-world datasets.

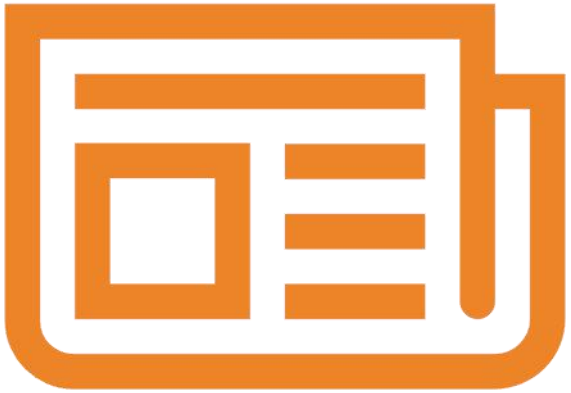
- Three driving factors for the Bike Demand:
  - Humidity levels
  - Temperature
  - Time of the year

- Accurate Rental Bike Demand Prediction

Introduction

- Analyze how weather would affect bike-sharing demand in urban areas
- Collect and process related weather and bike-sharing demand data from various sources
- Perform exploratory data analysis on the data
- Build predictive models to predict bike-sharing demand

Methodology



- Perform data collection
- Perform data wrangling
- Perform exploratory data analysis (EDA) using SQL and visualization
- Perform predictive analysis using regression models
  - Build the baseline model
  - Improve the baseline model
- Build a R Shiny dashboard app

Methodology

Data collection

Scrape Bike Sharing Table from Wikipedia into R

Data converted into .csv file

Setup OpenWeather API Account

Call Content into data frame

Store Tables of Raw Data As CSV Files

Data wrangling

Standardize Column Names

Read Data Sets to dataframes

Collect related variables for subset data frames

Determine the category of selected variables

Determine reference links in variables

Determine reference links in variables

Extract values using regular expressions

EDA with SQL

- Determine the observation count
- Determined the number of hour observations
- Queried weather forecast
- Determine the seasons which are included in the dataset
- Determined the highest bike rental counts in hours
- Determine the average value of weather conditions for each season on the dataset

EDA with data visualization

- Scatter plots describe the correlation between dates and bike rental counts
- Scatter plot, describing the correlation between dates, bike rental counts and hours.
- Density histogram of bicycle rental count, overlapped with the same density map.
- Multi-faceted scatter chart showing the correlation between time, bike hire count and temperature according to season.
- Scatter chart, which describes the relationship between working hours, number of rented bikes and temperature.
- The box chart shows the distribution of the number of bicycles rented, divided by hour and season.

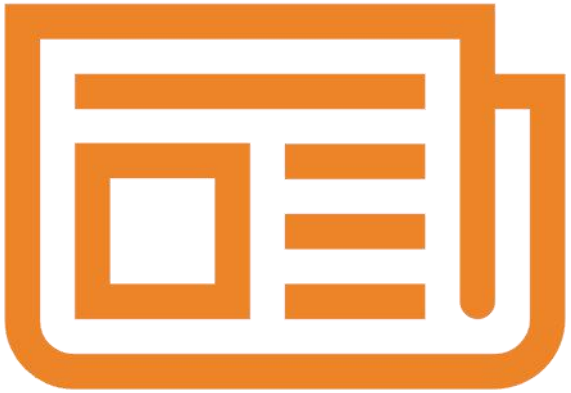
Predictive analysis

- Create a linear model (using weather variables only)
- Identify the most important variables
- Add polynomial terms, interaction terms, and regularization to the model
- Check RMSE and R-Squared on the model to find the best model

Build a R Shiny dashboard

- A global map depicting the level of prediction of sample cities
- Draw a 5-day temperature forecast
- Forecast bicycle demand by date
- Bicycle rental projections based on humidity levels

Results



- Exploratory data analysis results
- Predictive analysis results
- A dashboard demo in screenshots

EDA with SQL

Busiest bike rental times

- Find dates and hours which had the most bike rentals

- The highest number of bikes was recorded at the 18th hour on June 19, 2018, with a record value of 3,556 bikes per hour

A data.frame: 1 × 3

	DATE	HOUR	MAX_BIKE_COUNT
	<fct>	<int>	<int>
1	19/06/2018	18	3556

Rental Seasonality

- Rental Seasonality

- On average, the number of bikes is highest in summer, followed by demand in autumn. However, the higher standard deviation of the measured values suggests that, in each season, there are further determinants affecting overall bicycle demand in any given season.

A data.frame: 4 × 5

	AVERAGE_BIKE_COUNT	MINIMUM_BIKE_COUNT	MAXIMUM_BIKE_COUNT	STD_COUNT	SEASONS
	<int>	<int>	<int>	<dbl>	<fct>
1	924	2	3298	617.3885	Autumn
2	746	2	3251	618.5247	Spring
3	1034	9	3556	690.0884	Summer
4	225	3	937	150.3374	Winter

Weather Seasonality

- Weather Seasonality

- There is a strong correlation between temperature, humidity and wind speed and the number of bike rentals. The higher the value of both, the higher the demand for bike rentals seems to

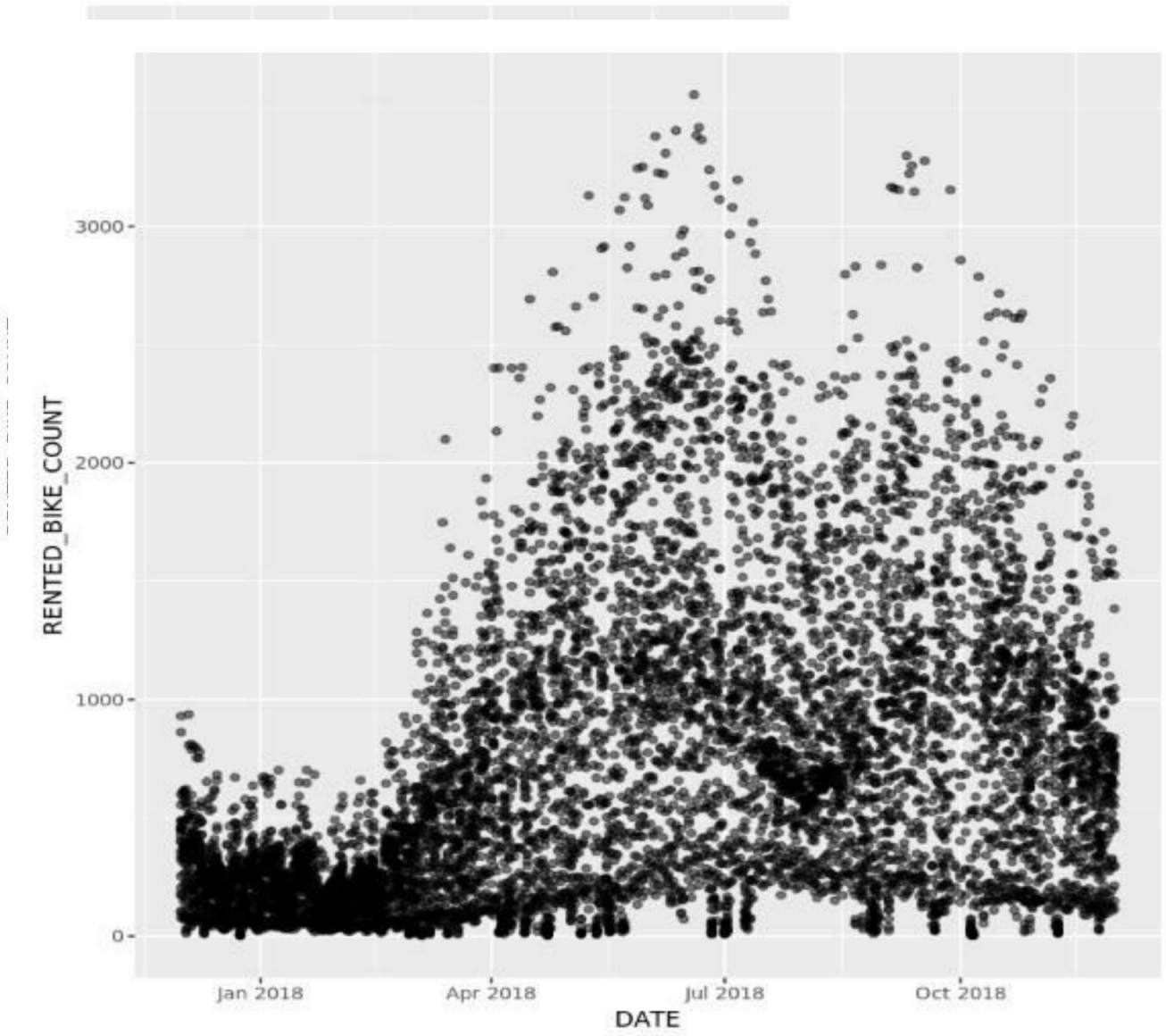
AVG_TEMPERATURE	AVG_HUMIDITY	AVG_WIND_SPEED	AVG_VISBILITY	AVG_DPT	AVG_SR	AVG_RAIN	AVG_SNOW	AVG_RENT_COUNT
<dbl>	<int>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
13.821167	59	1.492101	1558	5.150594	0.5227827	0.11765617	0.06350026	924
13.021389	58	1.857778	1240	4.091389	0.6803009	0.18694444	0.00000000	746
26.587274	64	1.609420	1501	18.750136	0.7612545	0.25348732	0.00000000	1034
-2.540463	49	1.922685	1445	-12.416667	0.2981806	0.03282407	0.24750000	225

EDA with Visualization

Bike rental vs. Date

Show a scatter plot
of RENTED_BIKE_COUNT vs. DATE

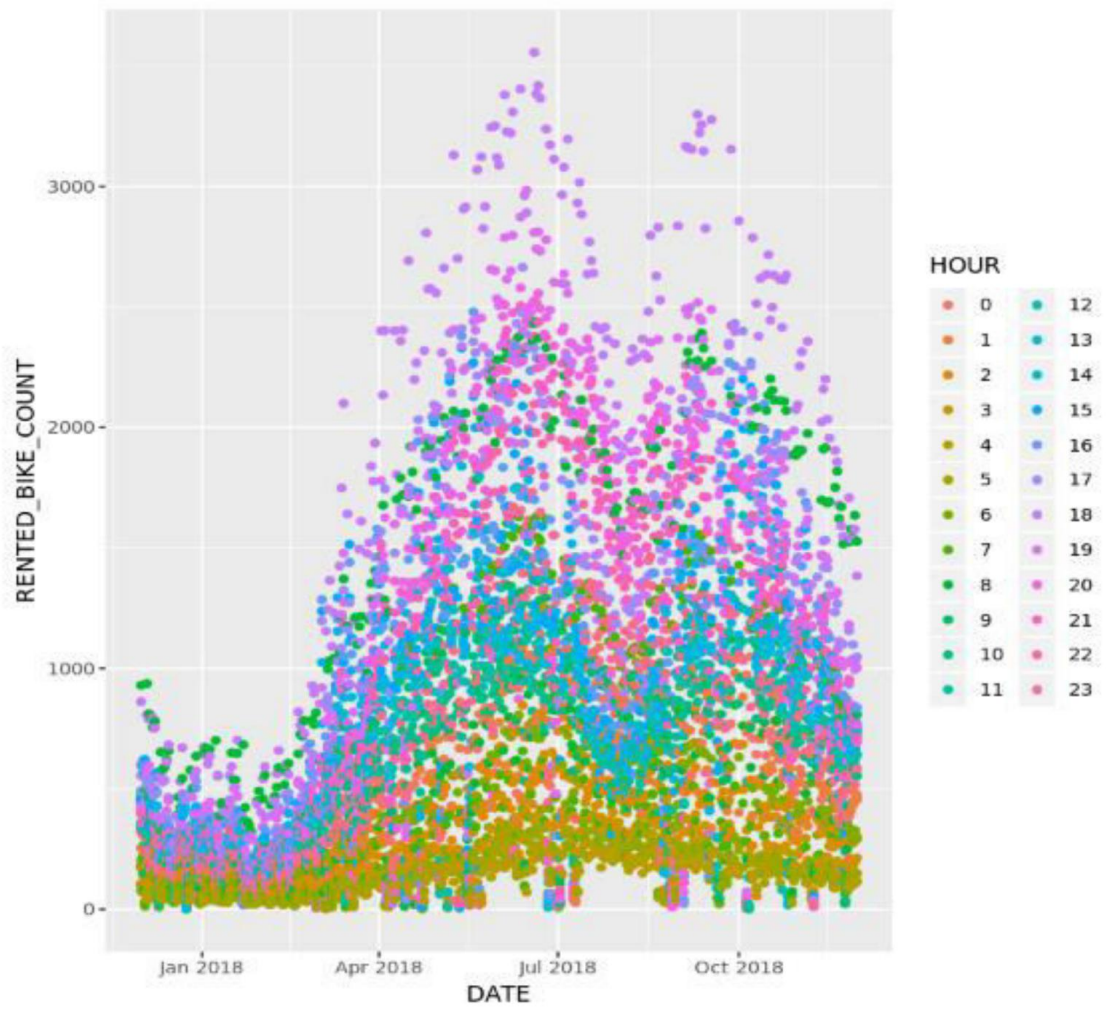
The scatter chart shows that, on average, the number of bike rentals is much higher in the summer and fall than at the end of winter. This pattern suggests that bicycle demand peaks in the middle of summer and towards the end of autumn.



Bike rental vs. Datetime

Show the same plot of the RENTED_BIKE_COUNT time series, but now add HOURS as the colour

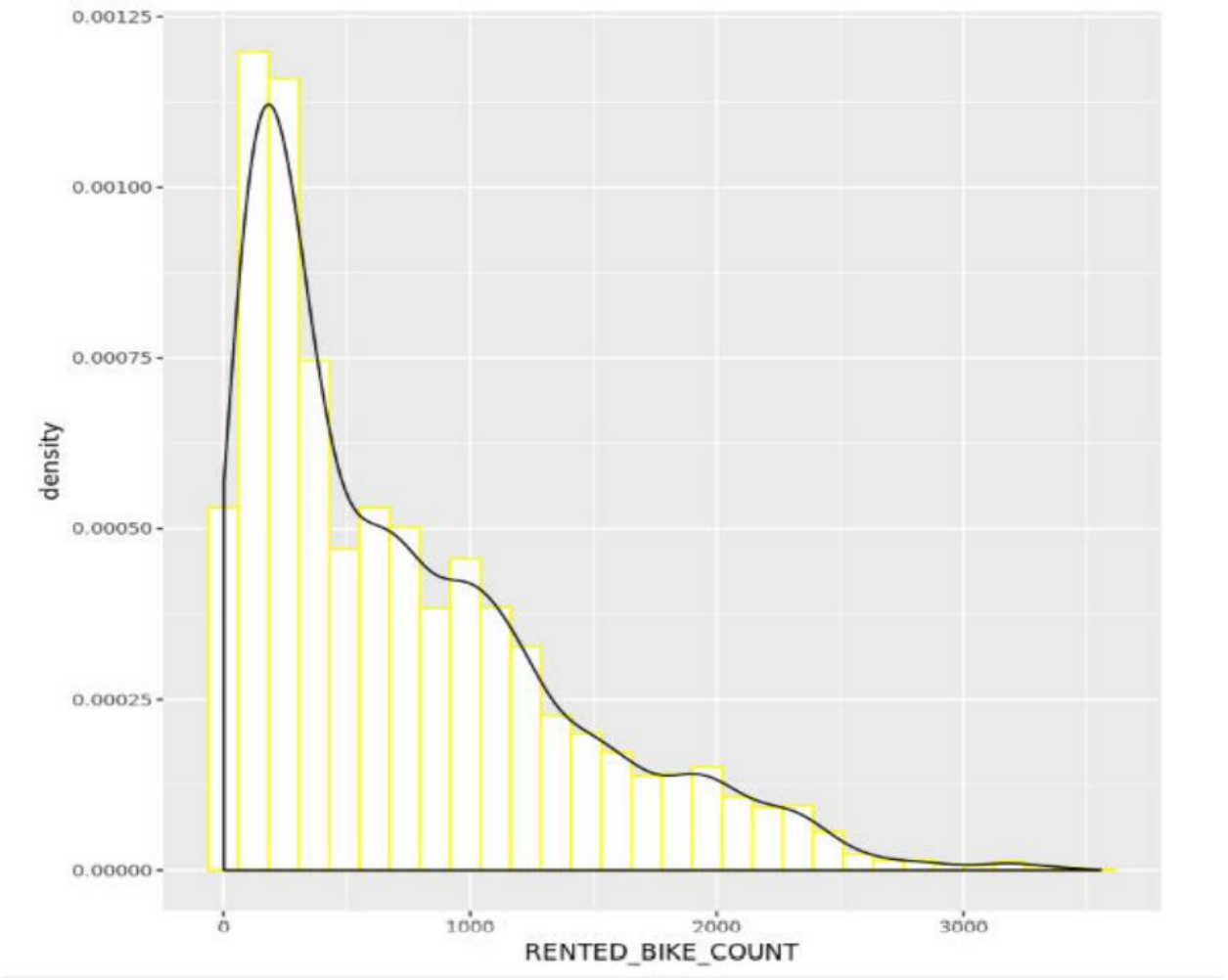
The scatterplot indicates that rental bike counts are much higher on average in the summer and autumn months compared to the late winter months. This pattern indicates that bike demand peaks in mid summer and near the end of autumn.



Bike rental histogram

Show a histogram overlaid with a kernel density curve

The highest demand density seems to be between 0 and 1000 for bike rentals. This shows that, on average, most observation points are located at the 0-1000 bike rental count level across the entire data set. Surges in demand are unusual. This has a significant impact on bicycle procurement and provides an indicator of the level of bicycle supply in the region.

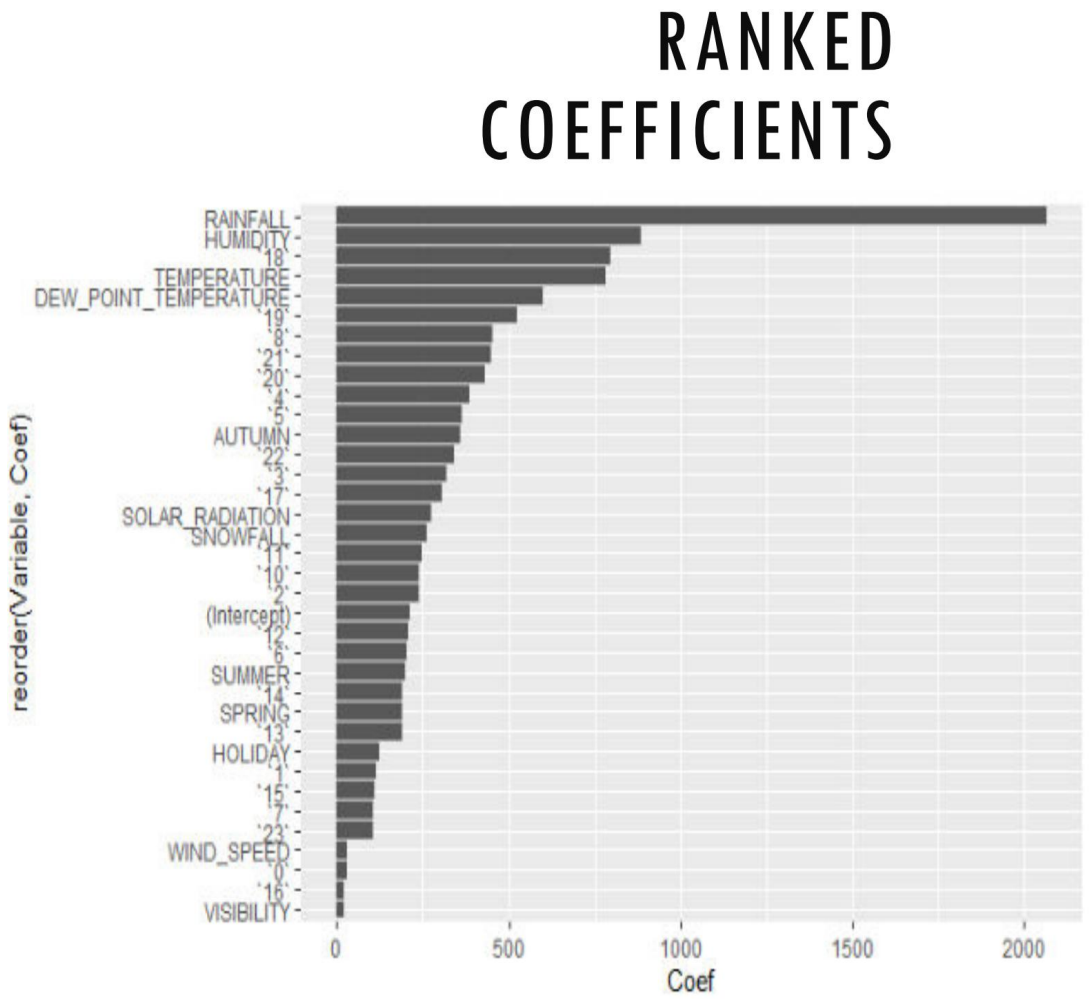


Predictive analysis

Ranked coefficients

Show a screenshot of the ranked coefficients bar chart for the baseline model

Given that the time of day is directly related to general weather conditions, the hour as a factor is also relatively important. Once you factor in rainfall, snow, temperature and humidity, seasonal variables become less important. After all, a season already contains these elements. Similarly, snow and solar radiation are less important factors, as bike demand is usually low in winter, and no one really looks for solar radiation values before deciding to leave home.



Model evaluation

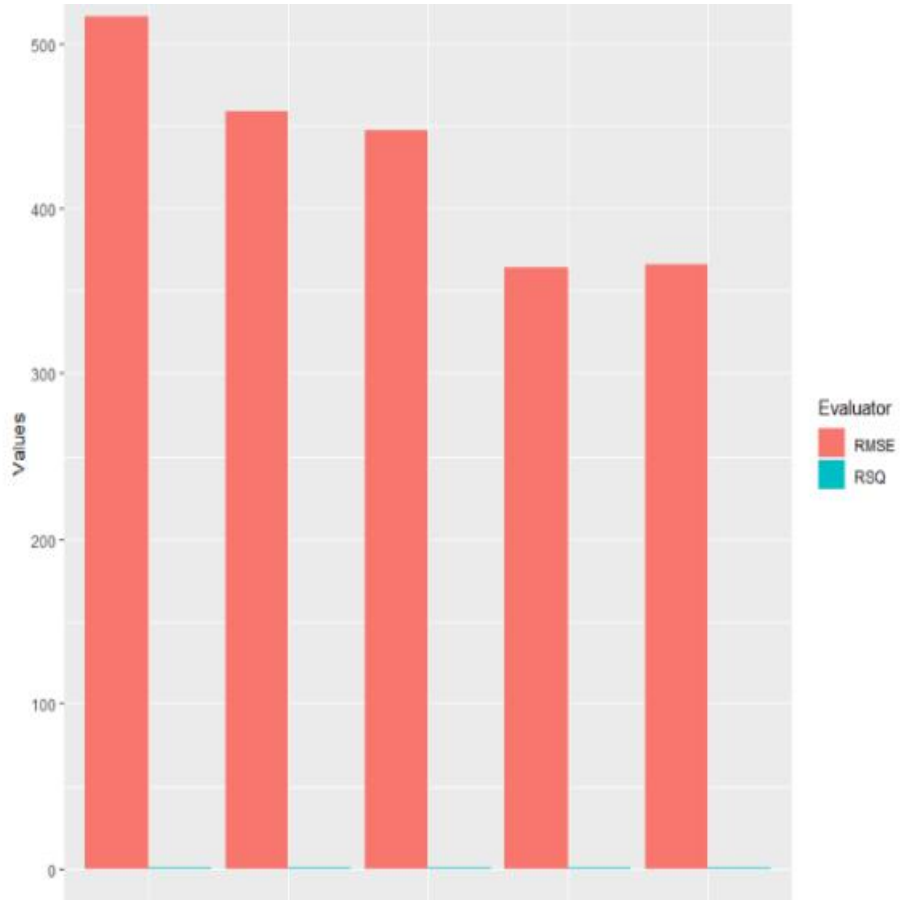
M1: Simple linear regression

M2: Quadratic polynomial regression

M3: Regression of cross-variables

M4: Return of lasso

M5: Ridge regression



Find the best performing model

- Show the model formula

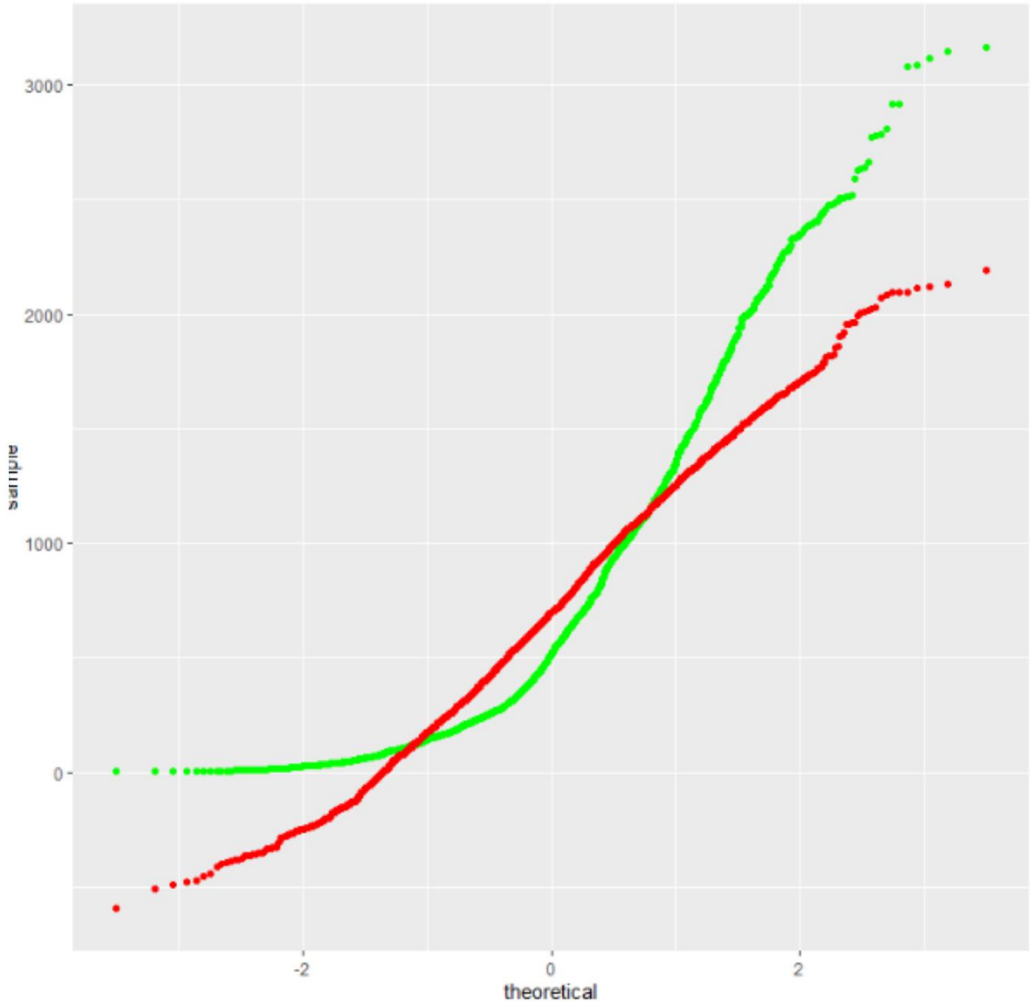
```
  best_model     RMSE       RSQ
1    model 4 364.6602 0.6685976
```

- The final coefficients

```
> lasso_fit%>%
+ extract_fit_parsnip()%>%
+ tidy()
# A tibble: 39 x 3
   term                  estimate penalty
   <chr>                    <dbl>   <dbl>
 1 (Intercept)              319.    0.036
 2 TEMPERATURE              966.    0.036
 3 HUMIDITY                -807.    0.036
 4 WIND_SPEED                30.7   0.036
 5 VISIBILITY                23.1   0.036
 6 DEW_POINT_TEMPERATURE    407.    0.036
 7 SOLAR_RADIATION          259.    0.036
 8 RAINFALL                -2074.   0.036
 9 SNOWFALL                 245.    0.036
10 0                         29.5   0.036
# ... with 29 more rows
```

Q-Q plot of the best model

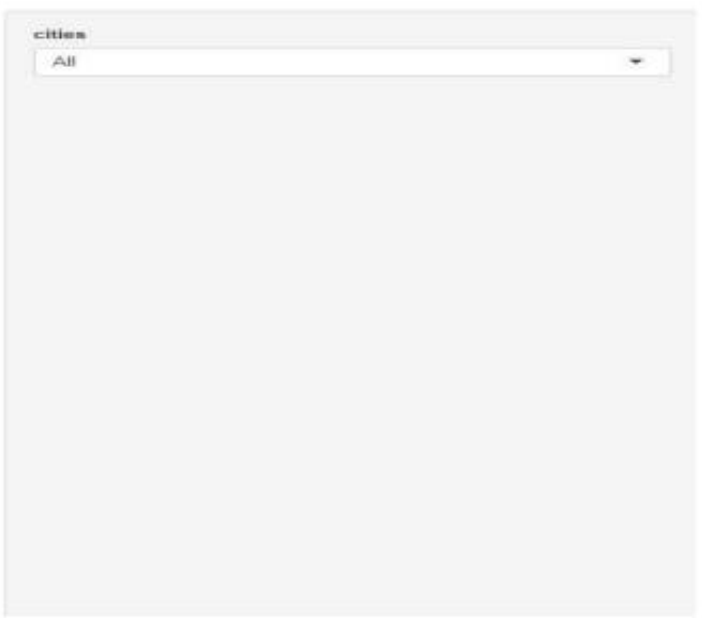
Plot the Q-Q plot of the best model's test results vs the truths



Dashboard

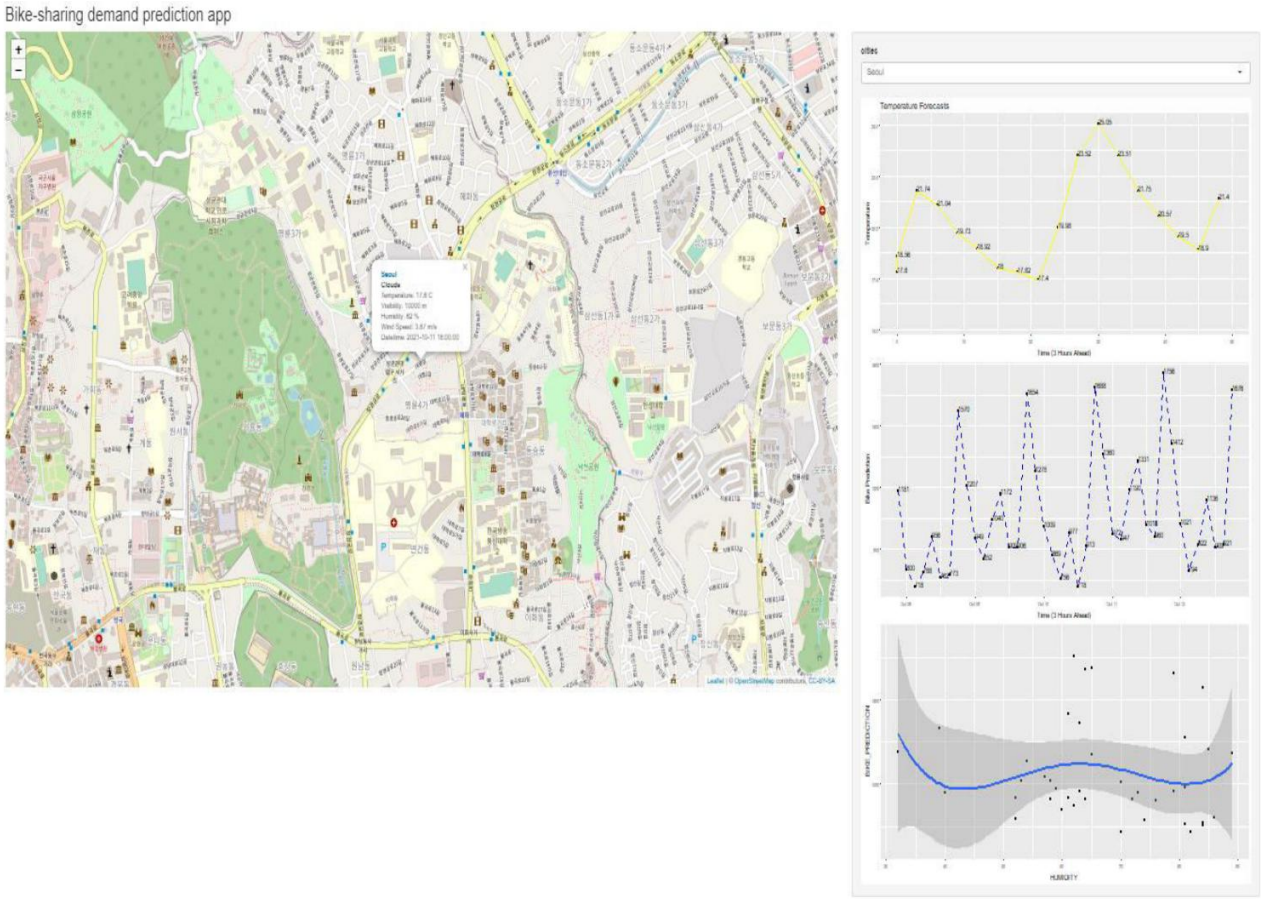
<GLOBAL MAX BIKE SHARING PREDICTION>

- The map above depicts the largest bike predictions for the five cities in our selected sample. A pop-up window describes the prediction level, where a unique color is specified as the bike prediction level associated with a given bike prediction value. In this case, the maximum bike prediction corresponds to the medium prediction level shown in a yellow pop-up box.



<DEMAND PREDICTIONS FOR SEOUL>

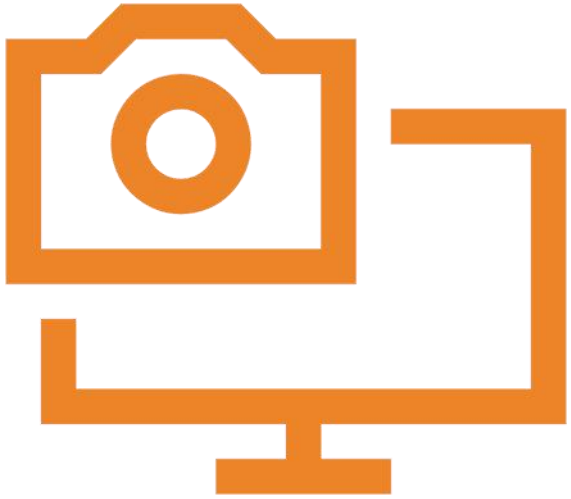
- The chart above shows current weather conditions in Seoul, and city map, temperature forecast for the next four days in the city. forecast bicycle demand during the selected time period, forecast bicycle demand based on humidity levels



CONCLUSION

- Bike rental demand is largely dependent on weather patterns.
- Temperature and humidity are important determinants of demand.
- Overall, demand for bike rentals is low among those selected.
- Accurate prediction models can improve supply chain decision-making.

APPENDIX



SQL Queries

```
query= "SELECT COUNT (RENTED_BIKE_COUNT) from SEOUL_BIKE_SHARING"
sqlQuery(Conn,query, believeNRows= FALSE)
```

```
query= "SELECT COUNT (HOUR) from SEOUL_BIKE_SHARING
WHERE
RENTED_BIKE_COUNT != 0"
sqlQuery(Conn,query)
```

```
query= "SELECT * from CITIES_WEATHER_FORECAST
Where
CITY= 'Seoul' Limit 1 "
sqlQuery(Conn,query)
```

```
query= "SELECT DISTINCT(SEASONS) from SEOUL_BIKE_SHARING"
sqlQuery(Conn,query, believeNRows=FALSE)
```

```
query= "SELECT DATE, HOUR, RENTED_BIKE_COUNT as max_bike_count from SEOUL_BIKE_SHARING
Where
RENTED_BIKE_COUNT= (SELECT max(RENTED_BIKE_COUNT) FROM SEOUL_BIKE_SHARING)"
sqlQuery(Conn,query, believeNRows=FALSE)
```

SQL Queries

```
query= 'select avg(RENTED_BIKE_COUNT) as average_bike_count, min(RENTED_BIKE_COUNT) as minimum_bike_count,
max(RENTED_BIKE_COUNT) as maximum_bike_count, STDDEV(RENTED_BIKE_COUNT) as STD_COUNT, SEASONS from SEOUL_BIKE_SHARING
GROUP BY SEASONS'
sqlQuery(Conn,query)
```

```
query= 'select avg(TEMPERATURE) as avg_temperature, avg(HUMIDITY) as avg_humidity,
avg(WIND_SPEED) as avg_wind_speed, avg(VISIBILITY) as avg_visbility, avg(DEW_POINT_TEMPERATURE) as avg_DPT,
avg(SOLAR_RADIATION) as avg_SR, avg(RAINFALL) as avg_RAIN, avg(SNOWFALL) as avg_SNOW,
avg(RENTED_BIKE_COUNT) as avg_rent_count, SEASONS from SEOUL_BIKE_SHARING
GROUP BY SEASONS
ORDER BY (select avg(RENTED_BIKE_COUNT) from SEOUL_BIKE_SHARING) DESC'
sqlQuery(Conn,query)
```

```
query= 'select CITY_ASCII, B.COUNTRY, LAT, LNG, POPULATION, BICYCLES from WORLD_CITIES A, BIKE_SHARING_SYSTEMS B
Where
A.CITY_ASCII =  B.CITY'
sqlQuery(Conn,query)
```

GGPLOT CODE SNIPPETS

```
# provide your solution here
ggplot(grouped_data, aes(x=DATE, y= RENTED_BIKE_COUNT))+geom_point(alpha=0.5)
```

```
# provide your solution here
ggplot(grouped_data, aes(x=DATE, y= RENTED_BIKE_COUNT, colour=HOUR))+geom_point()
```

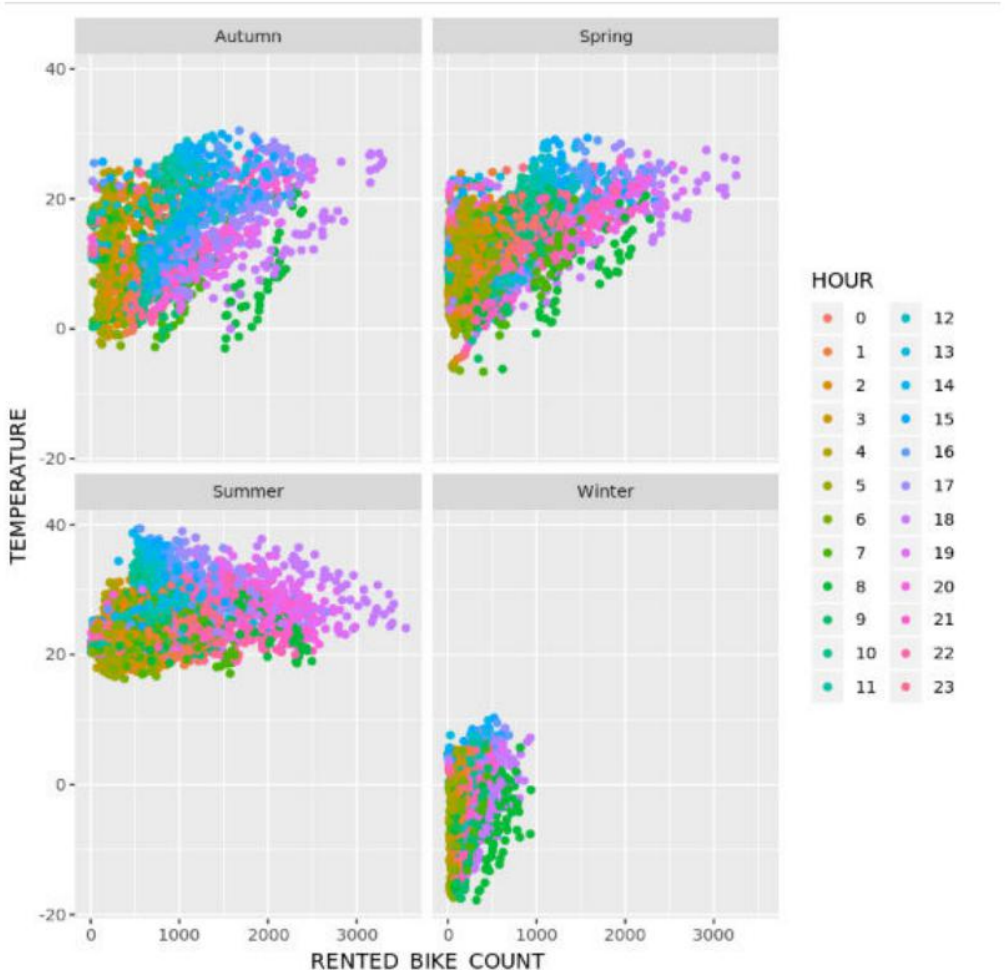
```
# provide your solution here
ggplot(grouped_data, aes(x= RENTED_BIKE_COUNT))+ geom_histogram(aes(y=stat(density)),fill="white", colour="yellow") + geom_densi
ty(colour="black")
```

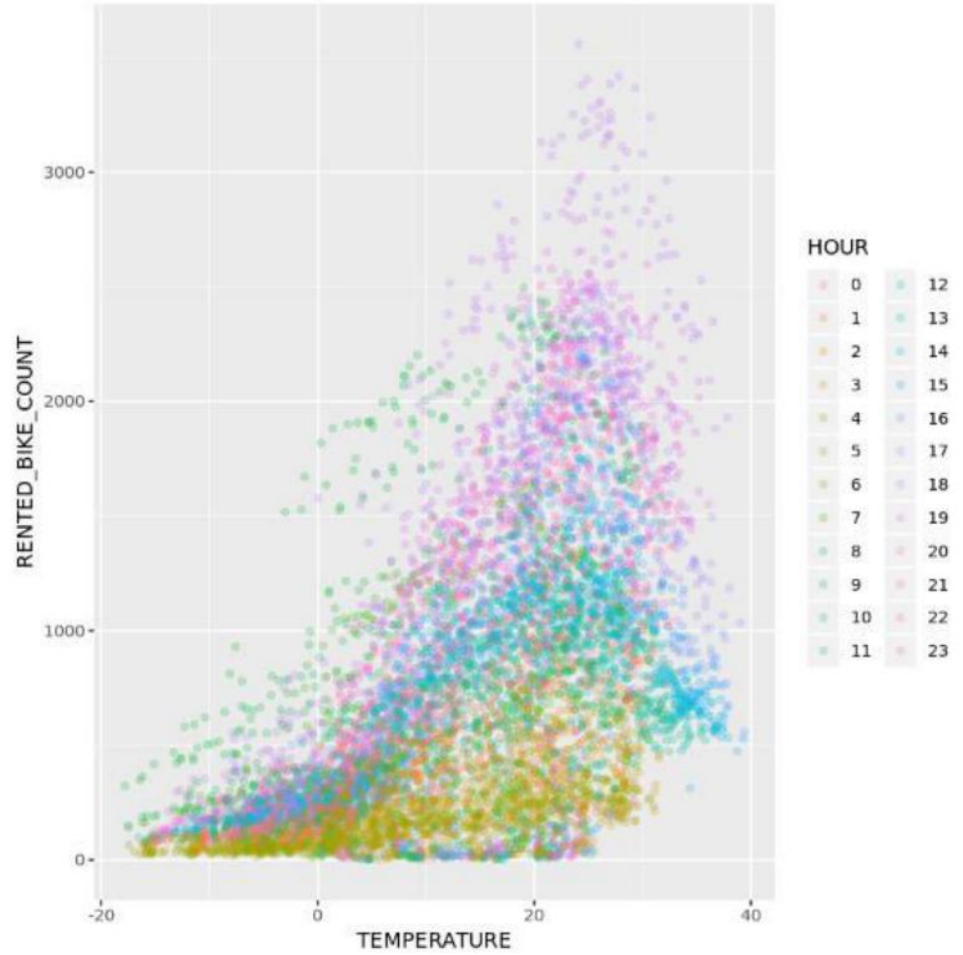
```
# provide your solution here
ggplot(grouped_data, aes(x=RENTED_BIKE_COUNT, y=TEMPERATURE))+geom_point(aes(color=HOUR))+facet_wrap(~SEASONS)
```

```
ggplot(grouped_data) +
   geom_point(aes(x=TEMPERATURE,y=RENTED_BIKE_COUNT,colour=HOUR),alpha=1/5)
```

```
# provide your solution here
ggplot(grouped_data, aes(RENTED_BIKE_COUNT, HOUR, group=SEASONS))+geom_boxplot()+facet_wrap(~SEASONS)+coord_flip()
```

PLOTS FOR FURTHER ANALYSIS





PLOTS CONTINUED