Project 4: Bike Sharing Demand

Forecast use of a city bikeshare system

Problem Statement

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these Bike Sharing systems, people rent a bike from one location and return it to a different or same place on need basis. People can rent a bike through membership (mostly regular users) or on demand basis (mostly casual users). This process is controlled by a network of automated kiosk across the city.

You are asked to forecast bike rental demand of a Bike sharing program in Washington, D.C. based on historical usage patterns in relation with weather, time and other data.

About The Data

The dataset shows hourly rental data for two years (2011 and 2012). The training data set is for the first 19 days of each month. The test dataset is from 20th day to month's end. Your job is to predict the total count of bikes rented during each hour covered by the test set.

In the training data set, they have separately given bike demand by registered, casual users and sum of both is given as count.

Training data set has 12 variables (see next slide) and Test has 9 (excluding registered, casual and count).

Data Fields

datetime - hourly date + timestamp
season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday - whether the day is considered a holiday
workingday - whether the day is neither a weekend nor holiday
weather - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp - temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed
casual - number of non-registered user rentals initiated
registered - number of registered user rentals initiated
count - number of total rentals

Hypothesis Generation

Before exploring the data, you should think about the problem, domain, and form a hypothesis. Doing so usually helps us form better features later on, which are not biased by the data available in the dataset. List a few of your hypothesis which could influence the demand of bikes. There's no right or wrong answers! For example, one might form the following hypothesis:

- Daily Trend: Registered users demand more bike on weekdays as compared to weekend or holiday.

Importing Data

Load the data into separate DataFrames. One for the test data and another for the train data.

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count	data_set
0	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0	3	13	16	train
1	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0	8	32	40	train
2	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0	5	27	32	train
3	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0	3	10	13	train
4	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0	0	1	1	train

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	data_set
0	2011-01-20 00:00:00	1	0	1	1	10.66	11.365	56	26.0027	test
1	2011-01-20 01:00:00	1	0	1	1	10.66	13.635	56	0.0000	test
2	2011-01-20 02:00:00	1	0	1	1	10.66	13.635	56	0.0000	test
3	2011-01-20 03:00:00	1	0	1	1	10.66	12.880	56	11.0014	test
4	2011-01-20 04:00:00	1	0	1	1	10.66	12.880	56	11.0014	test

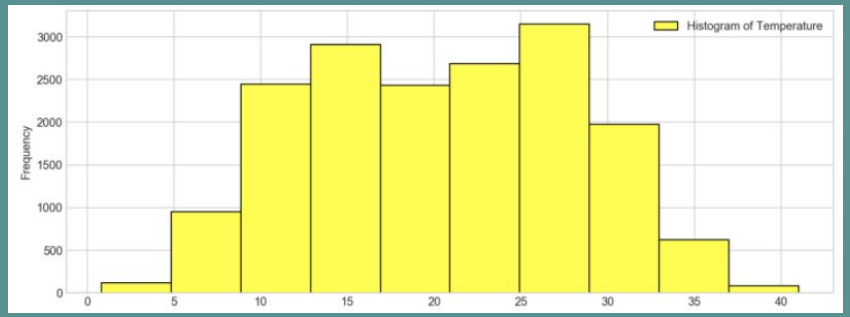
Data Exploration

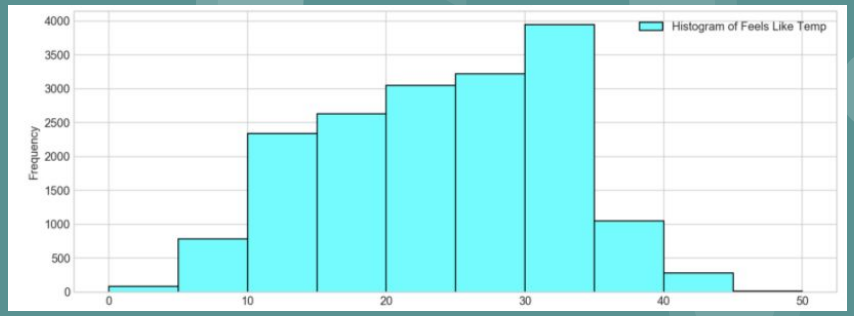
Combine both the train and test data set (to understand the distribution of independent variable together). That is, concatenate them into a new DataFrame that you'll use for data exploration. Hint: You'll have to set some default values to the columns not in the test data before concatenating the two together. The tail should looks something like below

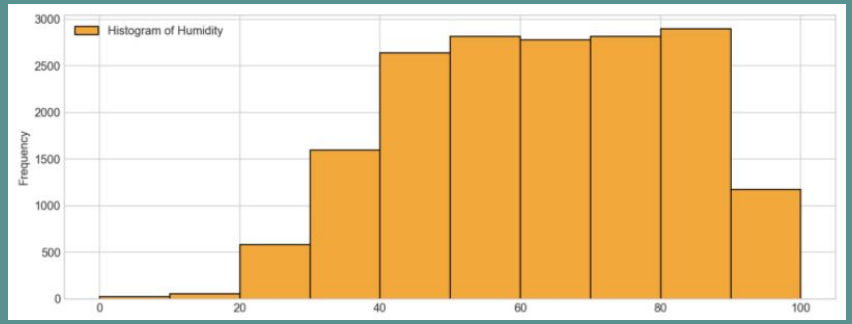
	atemp	casual	count	data_set	datetime	holiday	humidity	registered	season	temp	weather	windspeed	workingday
6488	12.880	0	0	test	2012-12-31 19:00:00	0	60	0	1	10.66	2	11.0014	1
6489	12.880	0	0	test	2012-12-31 20:00:00	0	60	0	1	10.66	2	11.0014	1
6490	12.880	0	0	test	2012-12-31 21:00:00	0	60	0	1	10.66	1	11.0014	1
6491	13.635	0	0	test	2012-12-31 22:00:00	0	56	0	1	10.66	1	8.9981	1
6492	13.635	0	0	test	2012-12-31 23:00:00	0	65	0	1	10.66	1	8.9981	1

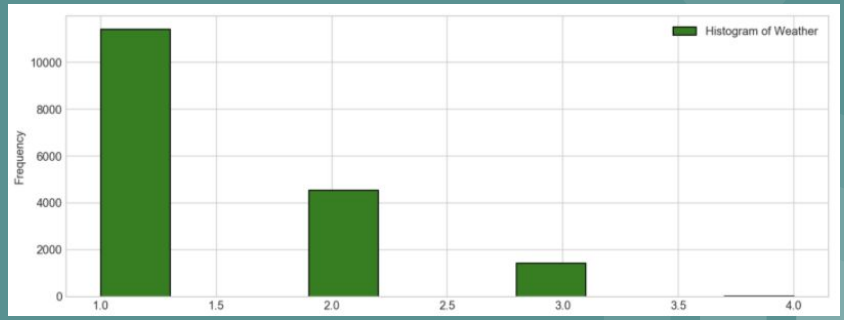
Data Exploration Cont...

Create a visualizations to help you understand the distribution of numerical variables and generate a frequency table for numeric variables. Below are some examples of what you should generate.







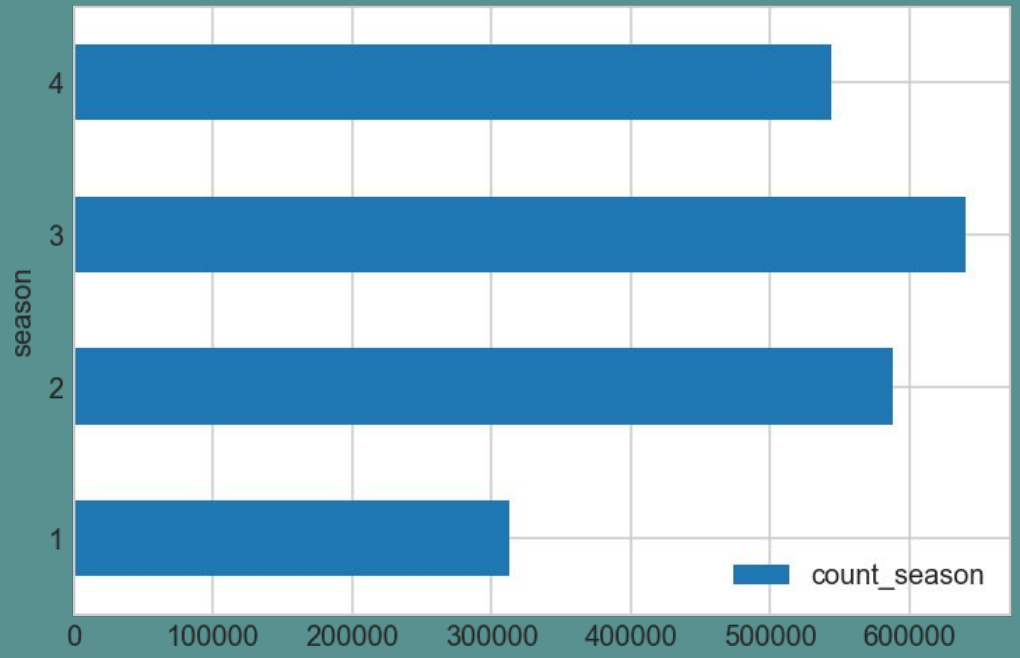


Data Exploration Cont...

List a few inferences that can be drawn by looking at the visualizations that you generated.

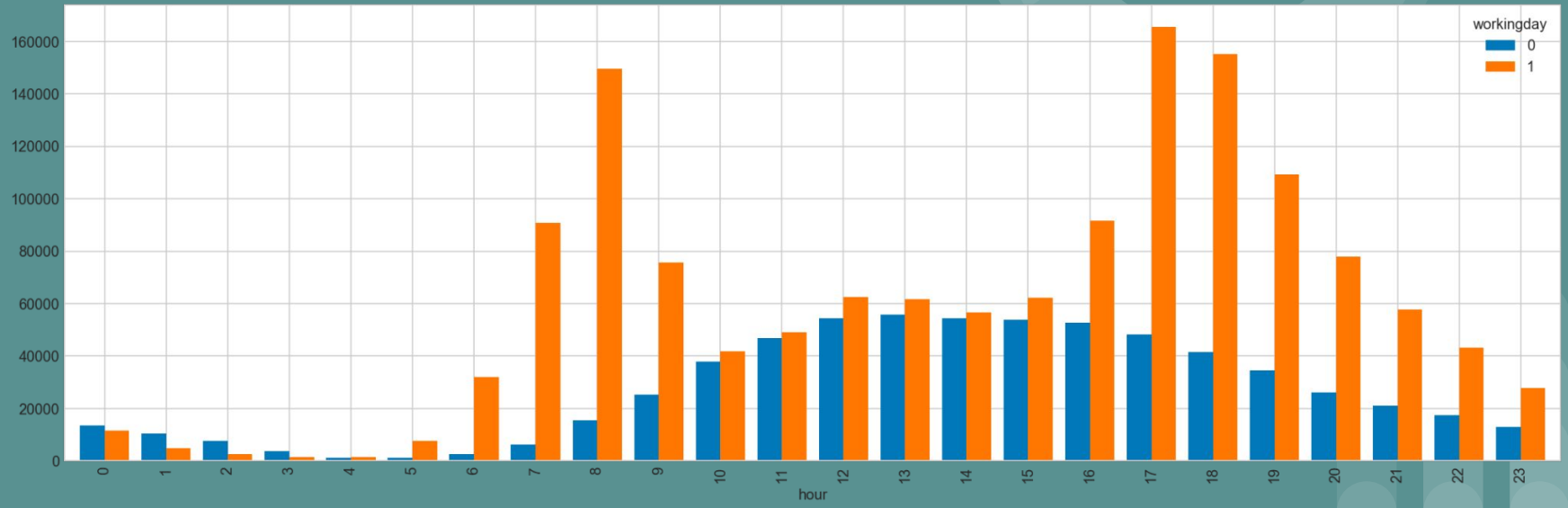
Data Exploration Cont...

Create a graph to visualize the season counts in the data set.



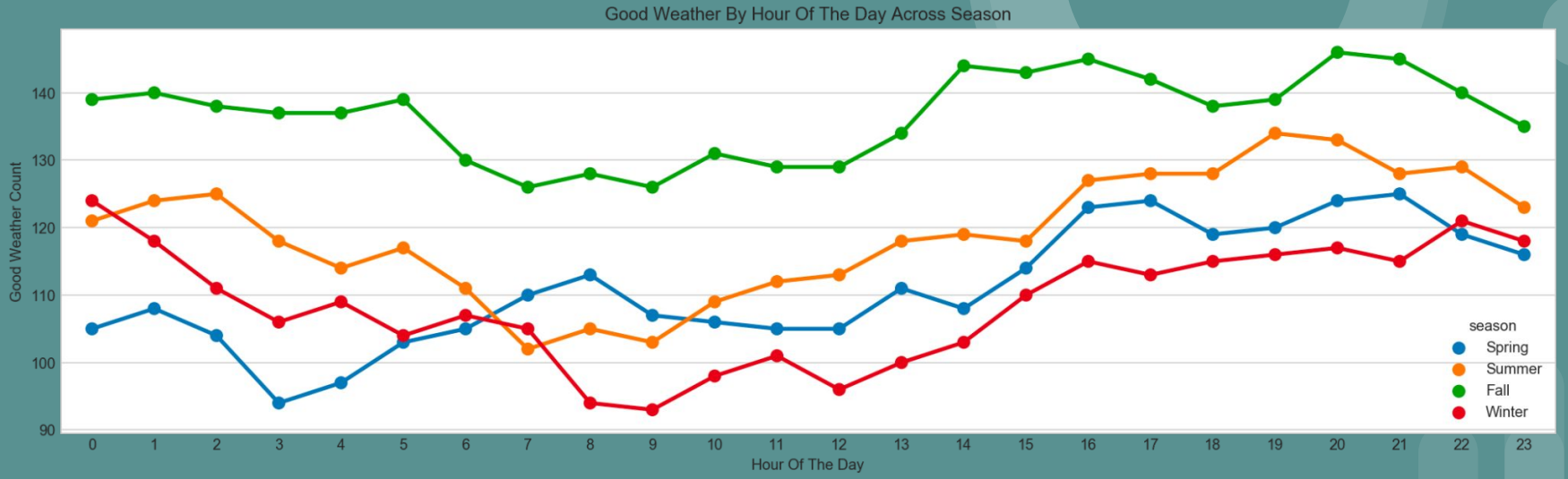
Data Exploration Cont...

Create some type of visualization(s) in order to inspect the hourly and workday trend. What can you conclude from your visualizations. Below is an example which you can try to recreate, but feel free to inspect this information in a different way if you'd prefer.



Data Exploration Cont...

Create a visualization to determine the following: Good Weather is most frequent in what season?



Data Exploration Cont...

Continue to explore the data on your own and keep track of any interesting conclusions. There's lots of possibilities to explore, blow are a few questions that you can ask yourself

- Which season is bad weather most frequent and typically during which hours?
- During which type of weather are bikes most rented in?
- How are renting patterns of bikes different between registered and casual users?
- Is there any correlation between features?
- What is the distribution of data between Train and Test set based on season, weather, etc.
- Is there any type of feature engineering that can be done?

Scaling

- Is it necessary to scale the data? What benefits would it provide?
- Which scaler will you use for this data set (if any), and why? Min Max, Standard, Robust, etc.

Preprocessing

- Is there any preprocessing that needs to be done? If so, which columns and why?

Model Training

For this portion, you are required to train and test 2 or more different models. For each model, you should list why you think this is worth trying and report its performance. If possible to answer with your approach, which features did you find to be most important for predicting? Is your approach parametric or non parametric?

Lastly, be sure to leave a conclusion on what you learned from the data, and why you believe each model scored the way that it did.