BIKE SHARING CASE STUDY

NAREN REDDY PALUPUNOORI

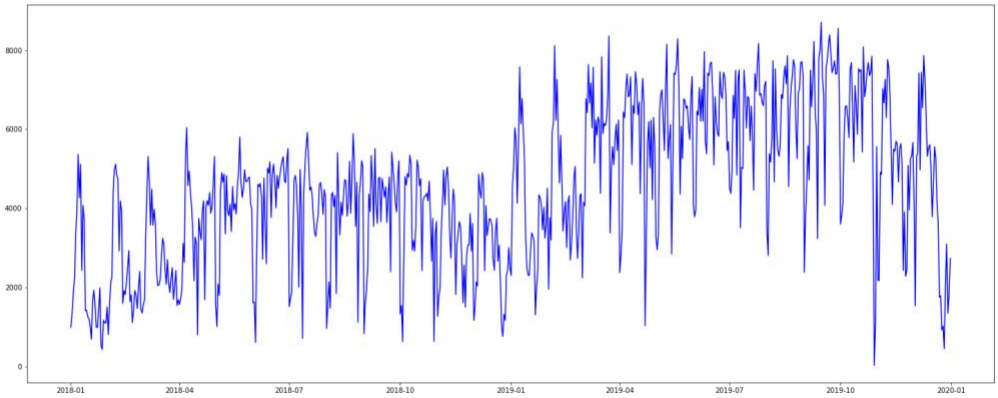
TABLE OF CONTENTS

Problem Statement

Understand the demand for bike sharing for given customers in the year of 2018 and 2019, to suggest marketing and predict demand for upcoming years post pandemic.

Understand variables effecting the bike demanding and how much these variables are affecting the demand.

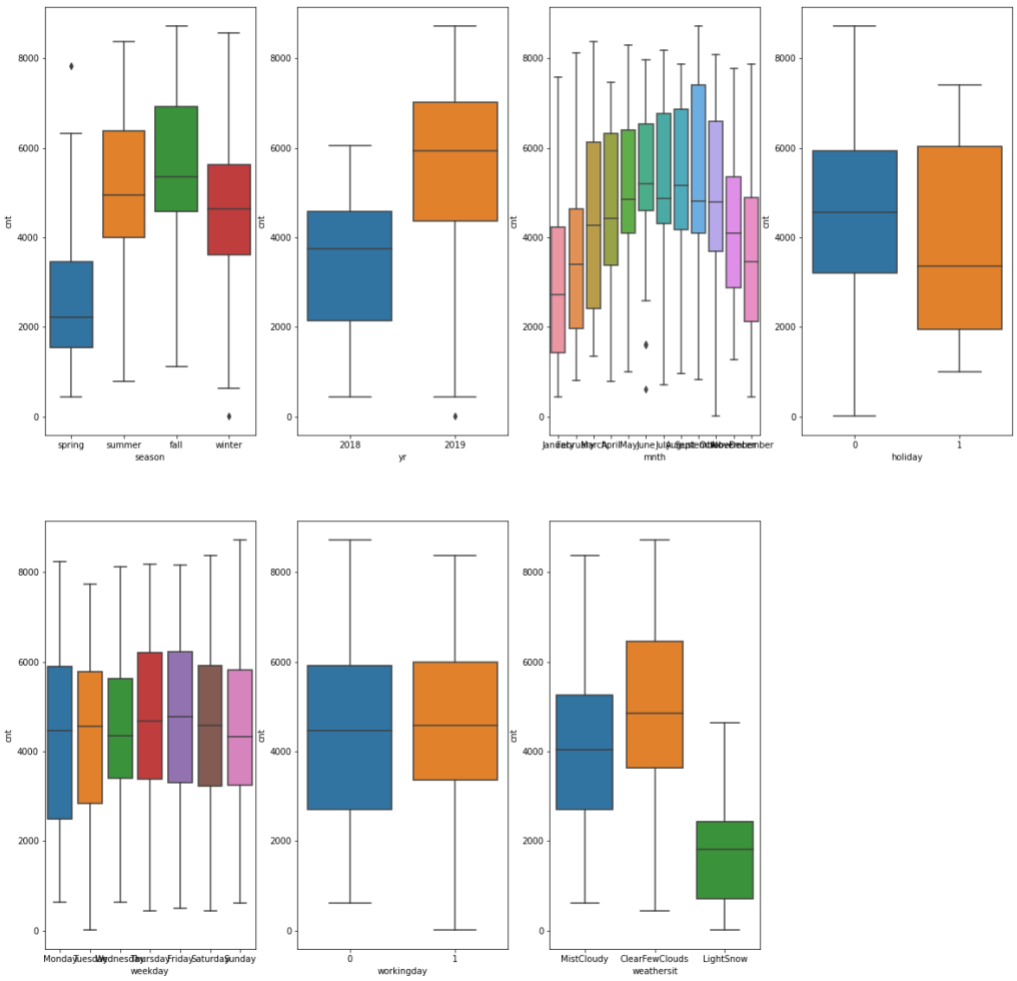
Analysis



- Demand for bike sharing has been steadily increasing from 2018 to 2019.

Relation analysis

- Season 1, spring has negative effect on bike sharing usage.
- Bike sharing has drastic uptick in 2019.
- Weekday and working day also have minimal effect.
- Snow has very drastic effect on bike sharing usage.

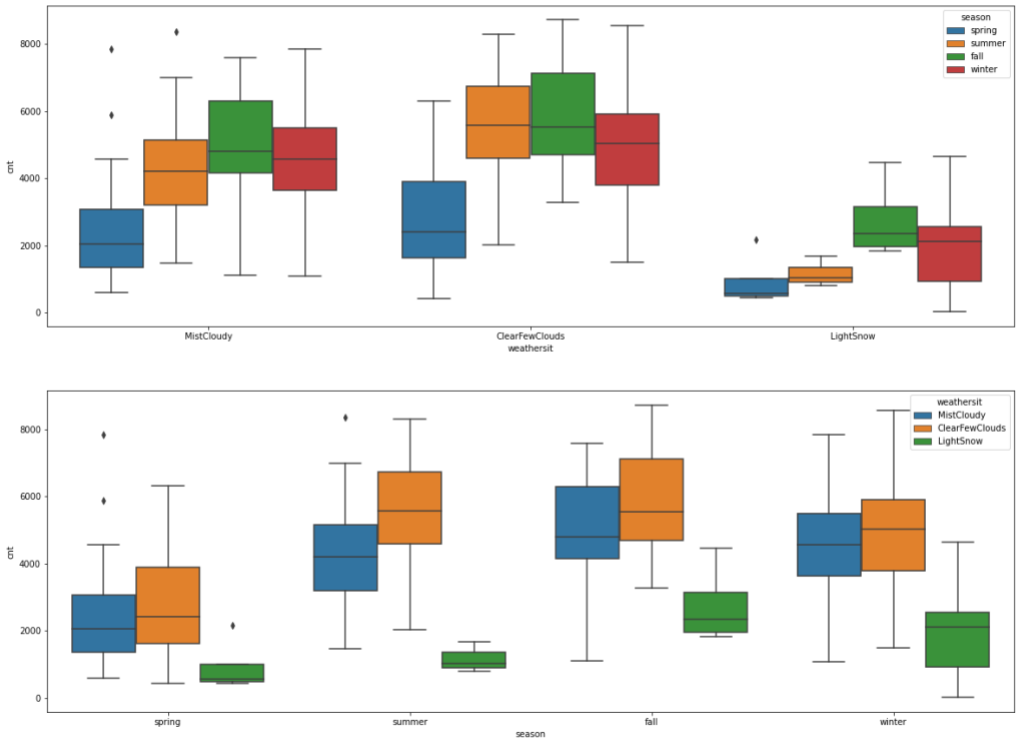


SEASON AND WEATHER IMPACT

There are two major observations from above graphs

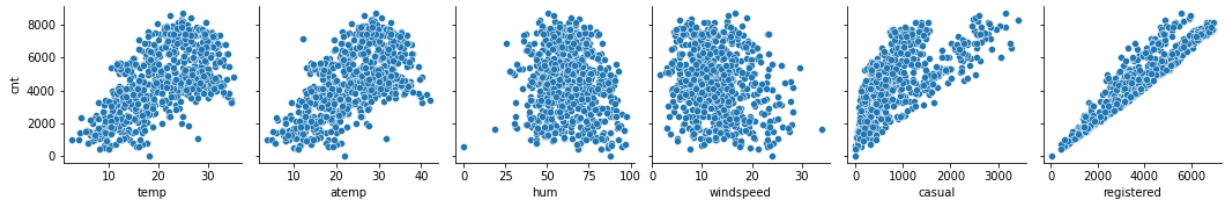
1. Season 1, spring has negative effect on bike sharing usage, irrespective of the weather.
2. Snow has a negative effect on bike sharing irrespective of the season.

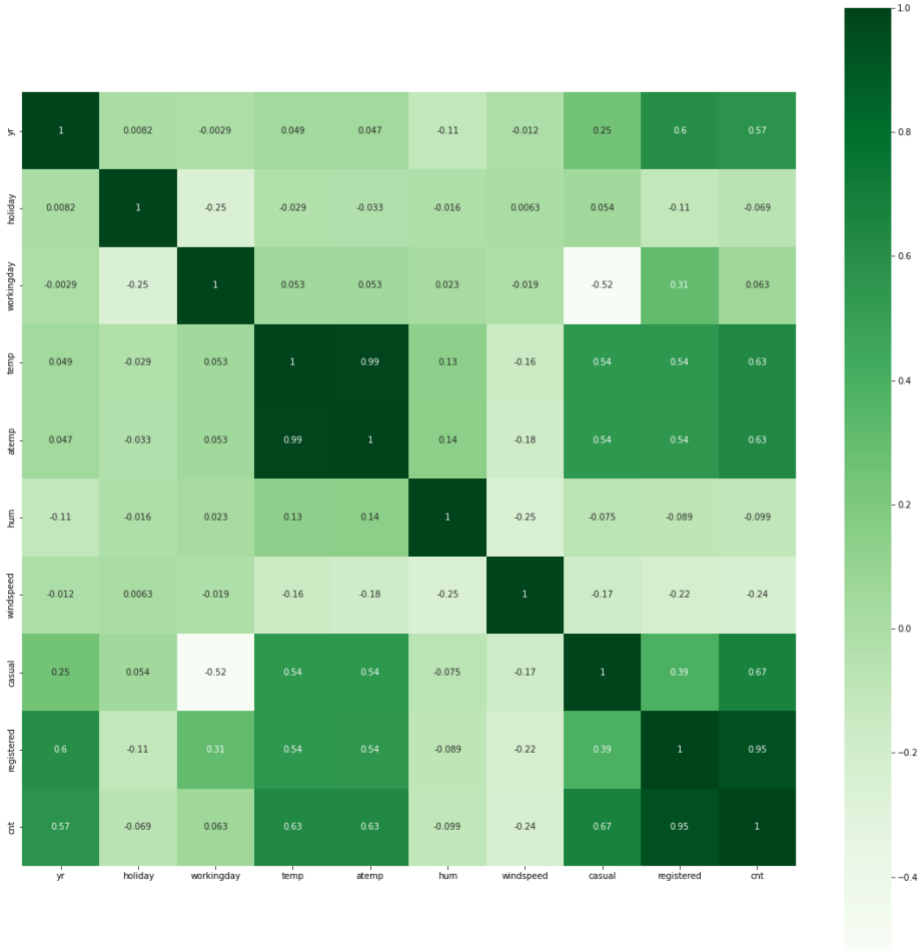
Conclusion here irrespective of seasons and weather, Snow and Spring has negative effect.



CORRELATION

- `cnt` is highly correlated to `registered` and `casual` which makes sense, `cnt` = `registered` + `casual`
- `atemp` and `temp` are highly dependent on the bike sharing. Understandable, based on temperature customers are deciding whether to use bike sharing or not.
- Also, there is some linear relation between `atemp`, `temp` and `cnt` so linear regression can be used here
- `season` and `weathersit` are also correlated to `cnt` as we saw `spring` and `snow` have high effect on customer use of bike sharing.




yr 1 0.0082 -0.0029 0.049 0.047 -0.11 -0.012 0.25 0.6 0.57
holiday 0.0082 1 -0.25 -0.029 -0.033 -0.016 0.0063 0.054 -0.11 -0.069
workingday -0.0029 -0.25 1 0.053 0.053 0.023 -0.019 -0.52 0.31 0.063
temp 0.049 -0.029 0.053 1 0.99 0.13 -0.16 0.54 0.54 0.63
atemp 0.047 -0.033 0.053 0.99 1 0.14 -0.18 0.54 0.54 0.63
hum -0.11 -0.016 0.023 0.13 0.14 1 -0.25 -0.075 -0.089 -0.099
windspeed -0.012 0.0063 -0.019 -0.16 -0.18 -0.25 1 -0.17 -0.22 -0.24
casual 0.25 0.054 -0.52 0.54 0.54 -0.075 -0.17 1 0.39 0.67
registered 0.6 -0.11 0.31 0.54 0.54 -0.089 -0.22 0.39 1 0.95
cnt 0.57 -0.069 0.063 0.63 0.63 -0.099 -0.24 0.67 0.95 1
yr holiday workingday temp atemp hum windspeed casual registered cnt
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

Features	cnt
cnt	1.000000
atemp	0.630685
yr_2019	0.569728
season_spring	0.561702
weathersit_LightSnow	0.240602
windspeed	0.235132
mnth_January	0.234235
weathersit_MistCloudy	0.170686
mnth_December	0.154565
mnth_February	0.150357
season_summer	0.145325
mnth_September	0.135876
mnth_June	0.119480
mnth_August	0.098590
hum	0.098543
mnth_July	0.096330
mnth_May	0.087758
holiday	0.068764
season_winter	0.064619
workingday	0.062542
mnth_October	0.058860
mnth_November	0.051795
weekday_Monday	0.037278
mnth_March	0.030720
weekday_Tuesday	0.026782
weekday_Saturday	0.023090
weekday_Wednesday	0.021187
weekday_Thursday	0.019302
weekday_Sunday	0.008087

MODEL TRAINING

	Features	CNT_enabled	CNT_Ranking	REG_enabled	REG_Ranking	CAS_enabled	CAS_Ranking
0	holiday	True	1	False	9	True	1
1	workingday	False	3	True	1	True	1
2	atemp	True	1	True	1	True	1
3	hum	True	1	True	1	True	1
4	windspeed	True	1	True	1	True	1
5	season_spring	True	1	True	1	True	1
6	season_summer	True	1	False	3	True	1
7	season_winter	True	1	True	1	False	14
8	weathersit_LightSnow	True	1	True	1	True	1
9	weathersit_MistCloudy	True	1	True	1	True	1
10	yr_2019	True	1	True	1	True	1
11	mnth_August	False	9	False	13	False	9
12	mnth_December	False	5	False	6	False	5
13	mnth_February	False	6	False	8	False	7
14	mnth_January	True	1	True	1	False	4
15	mnth_July	True	1	True	1	False	3
16	mnth_June	False	7	False	10	False	11
17	mnth_March	False	2	False	11	True	1
18	mnth_May	False	10	False	12	False	12
19	mnth_November	True	1	True	1	False	8
20	mnth_October	False	8	False	2	True	1
21	mnth_September	True	1	True	1	True	1
22	weekday_Monday	False	12	False	7	False	13
23	weekday_Saturday	True	1	False	4	True	1
24	weekday_Sunday	False	14	True	1	True	1
25	weekday_Thursday	False	13	False	5	False	2
26	weekday_Tuesday	False	11	False	14	False	6
27	weekday_Wednesday	False	4	True	1	False	10

Doing RFE model training as giving us a starting set of features. Analyzing the feature dependency,

- Holiday has very low impact on register customers but has a big effect on casual customers. As a result, there is some impact on overall demand.
- Even though working day does not have impact on overall demand, there is impact on registered and casual customers independently.

- During summer and winter there is no effect on demand, mostly due to casual customers.
- January and July there is demand change from registered customers no demand change from casual customers
- In March and October there is demand change for casual customers but no effect on overall demand.
- In November there is demand change form registered customers which is affecting overall demand.
- There is change in demand on Saturday from causal customers that is affecting overall demand.
- There is change in demand on Wednesday from registered customers, this is not affecting overall demand.

At this point of we can only analyze that there is change in demand. What the change is and how much is changed will be determined after doing the full model training.

Final Model analysis

Features	AllCoustomer	Registered	Casual
const	0.118783	0.059588	0.272966
atemp	0.604073	0.540362	0.514623
windspeed	-0.151717	-0.114052	-0.185519
season_summer	0.076528	NaN	0.082635
season_winter	0.119812	0.124169	NaN
weathersit_LightSnow	-0.255228	-0.262354	NaN
weathersit_MistCloudy	-0.074270	-0.066516	NaN
yr_2019	0.233042	0.253054	0.081171
mnth_September	0.073036	0.064347	NaN
workingday	NaN	0.140899	-0.234507
hum	NaN	NaN	-0.201971
mnth_October	NaN	NaN	0.055356
weekday_Sunday	NaN	NaN	0.041438

After final analysis there can some conclusions made and some suggestions to the marketing team

Conclusions

- With no other elements effecting the demand, there is more demand in casual customers then the registered customers.

- There is a positive demand increase in bikes for both registered and casual customers when it gets warmer.
    - For every 2 units increase in temperature there is 1 unit increase in demand.
- Demand for bike from casual customers tend to increase in summer. Increase in demand is very significant.
- Winter has more increase in demand then summer, all the demand increase is from registered customers.
- There is good increase in demand form registered customers in 2019, there is almost 25% increase in demand
- On windy days there is obvious drop in demand from all customers.
- There is loss of customers in snowy and cloudy days.
    - There is drop in demand from registered customers, Casual customers don't have much effect on demand during this season.
- Working day has opposite effect on demand from registered and casual customers, drop in demand from casual customers is compensated by demand from registered customers.
- Humidity as huge effect on casual customer's demand. Casual customers like to use bike's when the weather is better.
- During Sunday's there is uptick in demand from casual customers.

SUGGESTIONS

- Customer service team and supply team need be ready to handle demand increase in warmer days.
- Marketing team need to focus on getting more casual customers in summer and converting them to register customers, so they continue to become more regular users of bike and keep the demand up in winter also.
- In general, people are getting more health concussions this might contribute to increase in demand year on year.
- On working days looks like register customers are using bike sharing for commute to work. So, demand for bikes in residential area might increase during daytime and increase in demand in industrial area during evenings.
    - Marketing team need to focus on having some deals to get more casual customer to use during the evenings, like happy hour.
    - Loss of demand from casual customers is negating the demand increase from registered customers.
    - It is important to get more casual customers on working days.
- To handle the drop in demand in windy days, company need to think to upgrade to aerodynamic bikes and advertise it accordingly.

CONFIDENCE LEVEL

- From the model R^2 value we can be approximately 80% confidence on the suggestions and conclusions made for all customers and registered customers.

- For Casual customers we can be 70% confidence on the suggestions and conclusions.

	Customer type	R2 train data	R2 test data
0	All customers	0.816605	0.773717
1	Registered	0.829568	0.764625
2	Casual	0.705143	0.687274

FINAL EQUATIONS OF MODEL

$$cnt = 0.1188 + 0.6041 * atemp + (-0.2552) * weathersitLightSnow + 0.2330 * yr2019 + (-0.1517) * windspeed + 0.1198 * seasonWinter + 0.0765 * seasonSummer + (-0.0743) * weathersitMistCloudy + 0.0730 * mnthSeptember$$

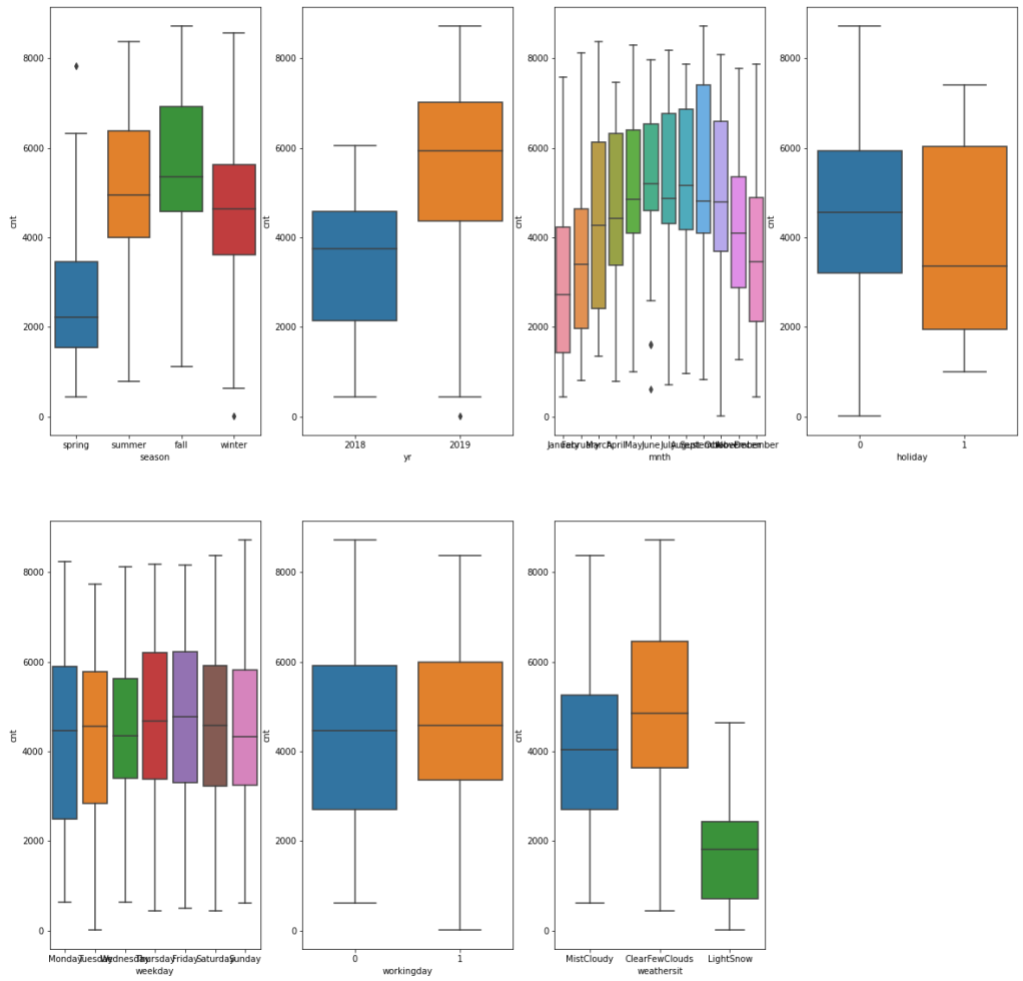
$$register = 0.0596 + 0.5404 * atemp + (-0.2624) * weathersitLightSnow + 0.2531 * yr2019 + 0.1409 * workingday + 0.1242 * seasonWinter + (-0.1141) * windspeed + (-0.0665) * weathersitMistCloudy + 0.0643 * mnthSeptember$$

$$casual = 0.2730 + 0.5146 * atemp + (-0.2345) * workingday + (-0.2020) * hum + (-0.1855) * windspeed + 0.0826 * seasonSummer + 0.0812 * yr2019 + 0.0554 * mnthOctober + 0.0414 * weekdaySunday$$

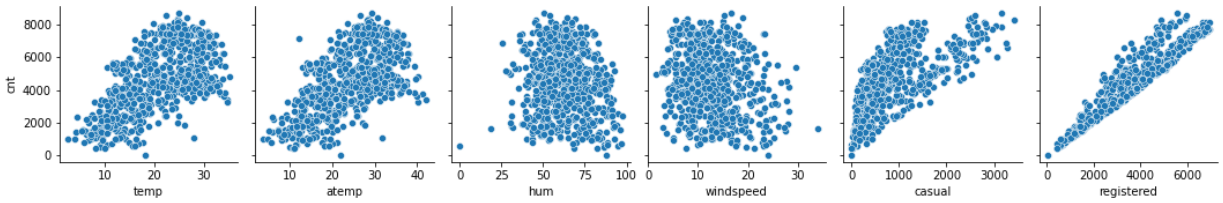
SUBJECTIVE

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
    a. Season 1, spring has negative effect on bike sharing usage.
    b. Bike sharing has drastic uptick in 2019.
    c. Weekday and working day also have minimal effect.
    d. Snow has very drastic effect on bike sharing usage.
    e. Conclusion here irrespective of seasons and weather, Snow and Spring has negative effect.

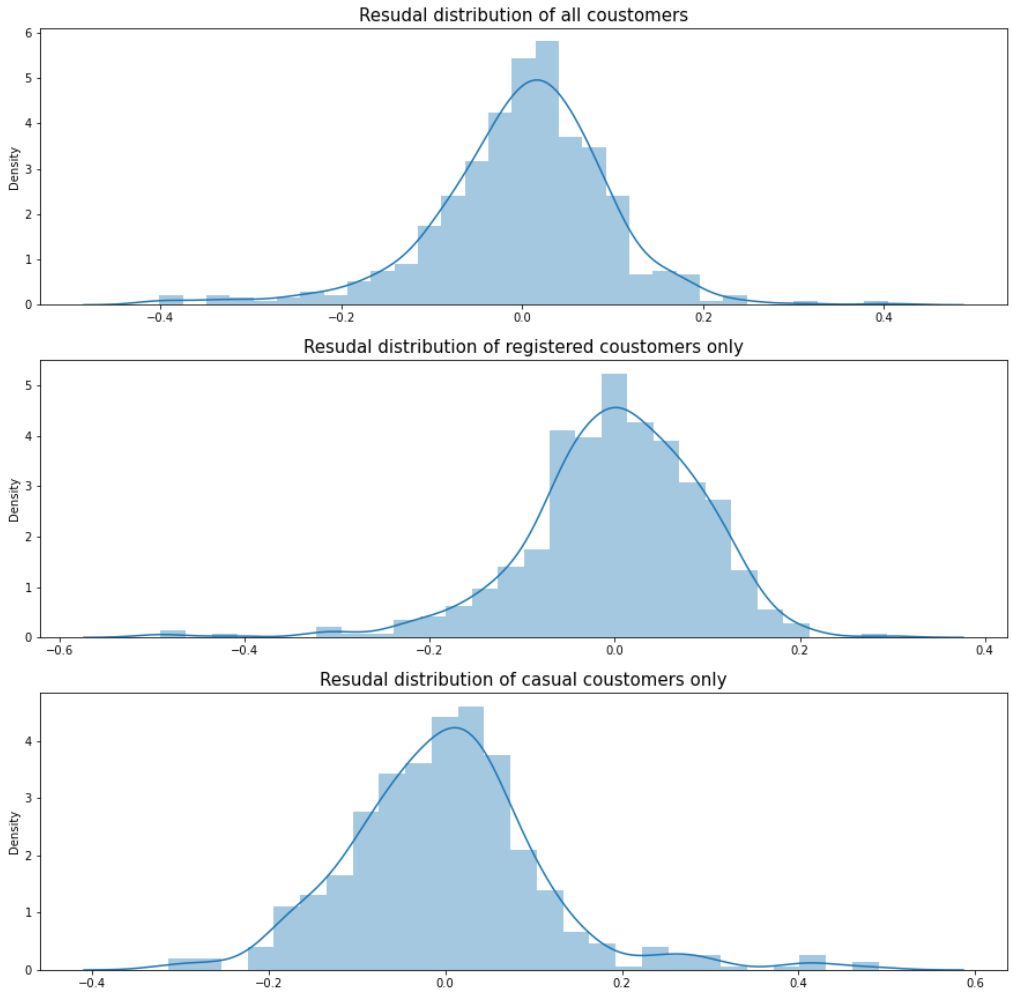


2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
    a. Drop first helps us to reduce the number of columns without losing actual content of the data.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    a. Registered has the highest correlation with the target variable cnt.
    b. Here since registered and casual columns are eliminated in the model, Next highest correlations ins with temp/atemp



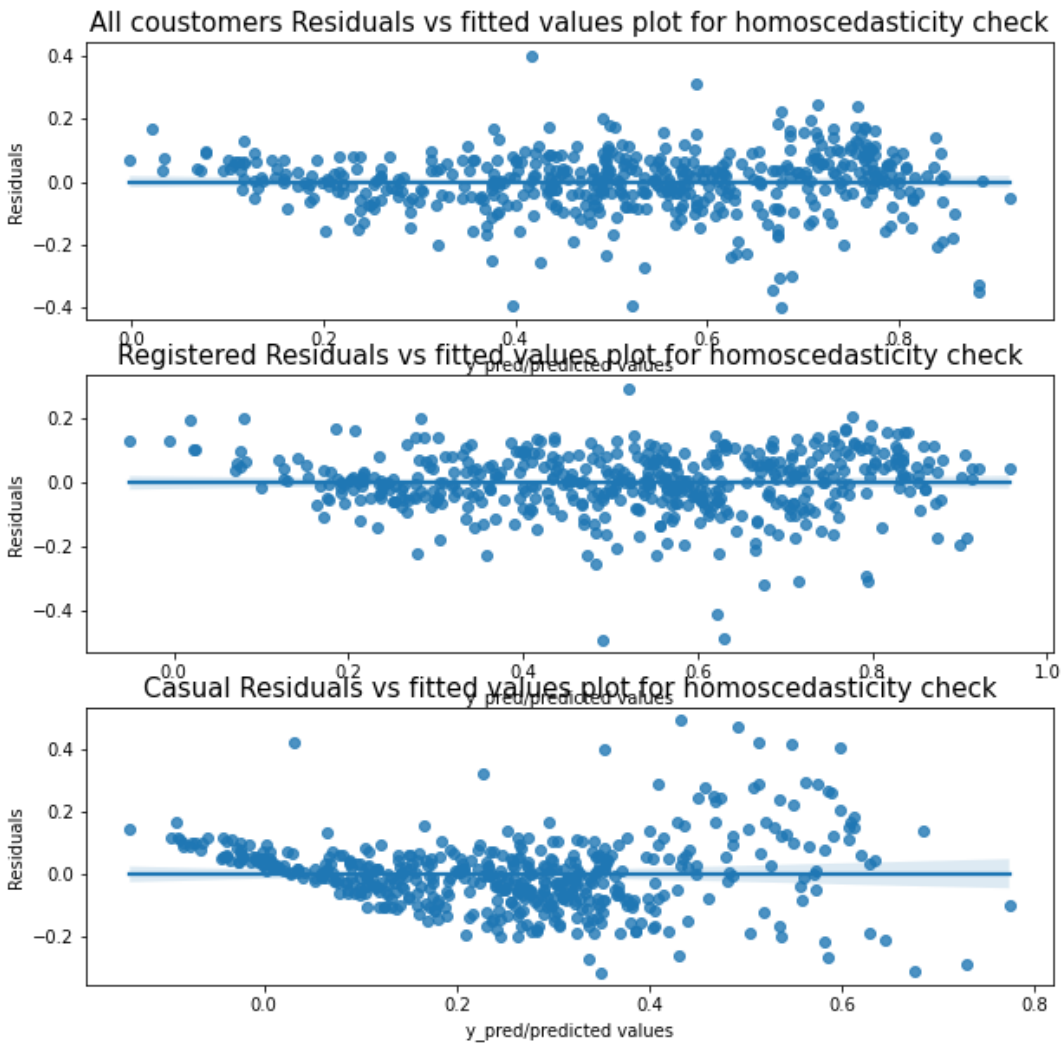
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. By doing residual analysis.
        i. Mean of residual analysis is zero

Stat	All coustomer	Registered	Casual_x	Casual_y
count	510.000	510.000	510.000	510.000
mean	-0.000	-0.000	-0.000	-0.000
std	0.097	0.095	0.111	0.111
min	-0.401	-0.492	-0.313	-0.313
25%	-0.046	-0.050	-0.066	-0.066
50%	0.008	0.004	-0.001	-0.001
75%	0.058	0.063	0.054	0.054
max	0.404	0.295	0.491	0.491



b. By checking for Homoscedasticity

Fitted line on residuals pass is almost flat with centered around 0, We can say residuals have equal variance. Another assumption of linear regression is satisfied



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Temperature and Temperature feel
    b. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    c. Year 2019
    d. Wind speed

Features	AllCoustomer	Registered	Casual
const	0.118783	0.059588	0.272966
atemp	0.604073	0.540362	0.514623
windspeed	-0.151717	-0.114052	-0.185519
season_summer	0.076528	NaN	0.082635
season_winter	0.119812	0.124169	NaN
weathersit_LightSnow	-0.255228	-0.262354	NaN
weathersit_MistCloudy	-0.074270	-0.066516	NaN
yr_2019	0.233042	0.253054	0.081171
mnth_September	0.073036	0.064347	NaN
workingday	NaN	0.140899	-0.234507
hum	NaN	NaN	-0.201971
mnth_October	NaN	NaN	0.055356
weekday_Sunday	NaN	NaN	0.041438

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. Linear regression algorithm is supervised algorithm explaining a liner relation between independent variable and dependent variable.
    b. Linear regression is defined by linear eq $y = c + a * x$
        i. Where c is the constant, intercept, value of dependent variable when there is no independent variable.
        ii. Y is the dependent variable.
        iii. X is the independent variable.
        iv. A is the slope of the equation explaining the rate of change of dependent variable when one unit of independent variable is changed.
2. Explain the Anscombe's quartet in detail.
    a. Anscombe's quartet is where the data distribution is not linear but have the same statical properties like,
        i. Same mean
        ii. Same variance
        iii. Same Min, Max
        iv. Same Mode

b. But when a line is drawn or residual analysis of linear model is done on this data, they fail to meet the linear regression assumption.

3. What is Pearson's R?
    a. Pearson correlation coefficient is a measure of linear correlation between two sets of data.
    b. Pearson coefficient is calculated as $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X * \sigma_Y}$
        i. cov is the covariance
        ii. σ_X is standard deviation of X
        iii. σ_Y us standard deviation of Y
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    a. Scaling is process of scaling all the values of a data set to same scale.
    b. There might be data with large value but very low correlation with dependent variable, misguiding the beta values, to avoid this we want all the features on same scale.
    c. In normalized scaling we map the feature values between 0 and 1.
        i. In normalized data there is a worry of the we might lose data spread since all the data is getting compressed between 0 and 1.
        ii. Outliers might be difficult to identify.
    d. In standardized scaling the data to have mean of 0 and standard deviation of 1.
        i. Here data still maintains the same spread of data.
        ii. Easy to identify outliers here.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
    a. VIF infinite between 2 variable means, these 2 variables can be perfectly explained with each other.
    b. 2 variables can be explained exactly by a linear combination.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
    a. Plot to graphically determine if 2 data sets came from the population with common distribution.
    b. Advantages:
        i. Sample size do not need to be equal
        ii. Distributions of 2 data sets can be handled at the same time as outliers.
    c. Need in linear regression:
        i. Helps to know if the test data or the data samples come from the population with same distribution.
        ii. If they are not same helps in understanding how they are different.