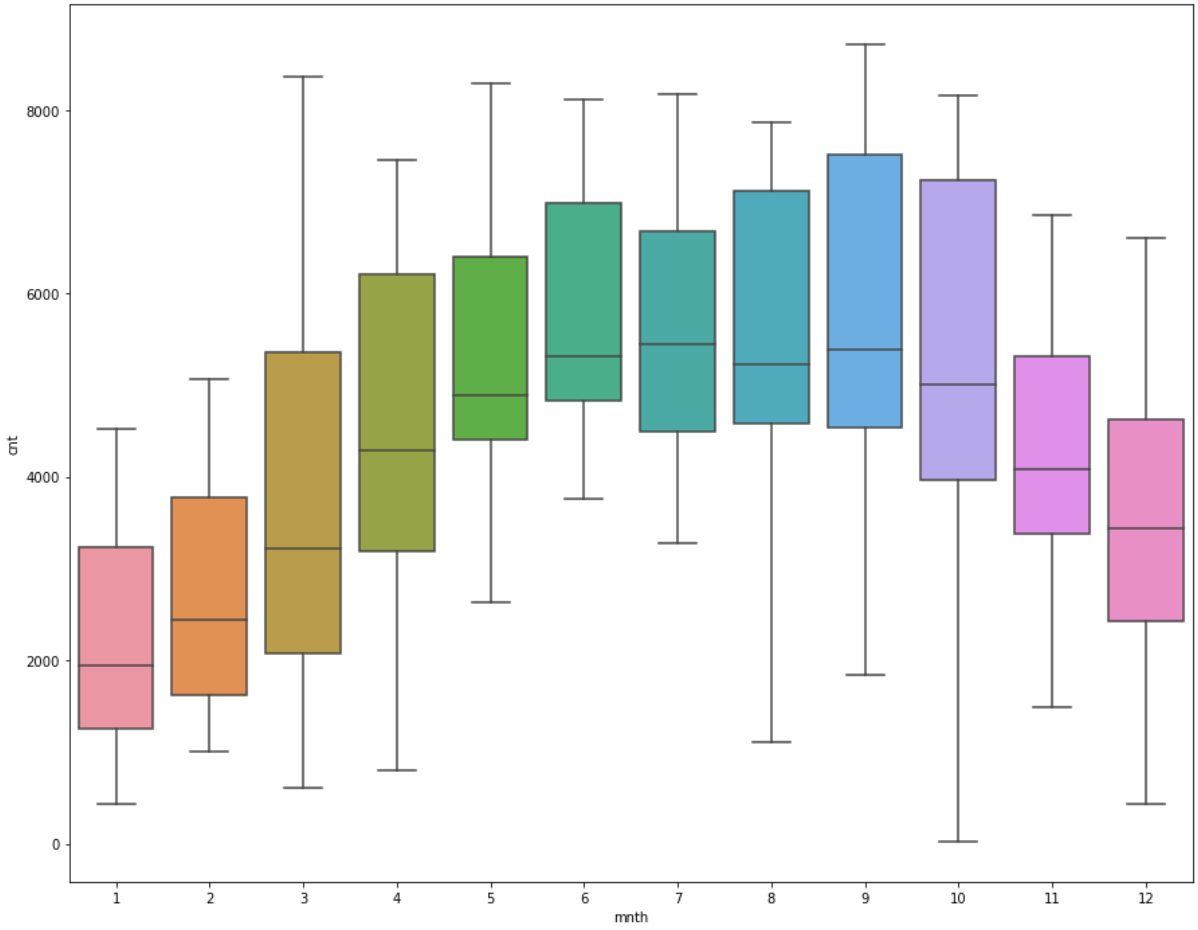
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Month, Season, Weathersit : Better Weather correlates to higher bike rentals



Some have little to no effect e.g. : Weekday, workingday

Holiday does show some effect but very few holidays in data so hard to say how relevant.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

By default, get_dummies creates dummies 1 for each unique value. But we need on n-1 for n, so we drop the first.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Temp, atemp both have R 0.65

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Checked the linearity of relationships between cnt and all the other vars.

Checked the Normal distribution of Residuals, using Histogram, Scatter plot & QQ Plot.

Errors should be normally distributed with mean of zero and same variance across the range of predicted values.

QQ plot points should hug the line.

In my case both the QQ plot & scatter plot suggest variance is not constant across the range of predicted values.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1. **temp : Postive**
2. **weathersit_3 : Negative**
3. **yr: Positive**

Good temperature make it easier to ride and drive higher rentals.

Weather situation 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. That is poor weather leads to lower rentals.

Year 2019 it seems had higher bike rentals.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

1. **Load the data and perform EDA**
2. **Split the data in Test and Training sets**
3. **Scale the Training data, (use this same scaling formula later on test)**
4. **If multiple features use RFE to reduce features to manageable number say 10-15**
5. **Look at VIF and remove multi-collinearity (5 or lower is generally acceptable) by removing variables one at a time**
6. **Iteratively add or remove the features till you have all significant coefficients, and good R square, qdj. R sq.**
7. **use the model on Test data and check R square**
8. **Check the model assumptions using residuals, normality, QQ Plot, scatter plot etc**

2. Explain the Anscombe's quartet in detail. (3 marks)

It is a data set with 4 categories, when looked just in terms of the simple stats. Mean/Std. dev.it seems that they are alike but plotting them shows that each is different.

https://learn.upgrad.com/course/3090/segment/20948/128448/393411/2046654

3. What is Pearson's R? (3 marks)

R is the correlation coefficient for 2 quantities/variables that shows the degree to which they vary together. It ranges from -1 to 1

-1 : perfect inverse relationship

0 : no relationship

1: perfect positive relationship

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is the transformation of continuous/numeric variables to bring them to common level/range. So that when model is built the coefficients get scaled to their importance/impact rather than the nominal values the variable takes.

 # Min Max Scaling (Normalisation): (x - xmin)/(xmax - xmin)

 # Standardization: (x - x^)/s.d.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

This will happen when at least 2 of the independent variables have correlation approaching 1. E.g in our case atemp & temp had nearly R = 1 . So we may assume some of the independent vars are perfectly correlated.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q here stands for Quantile- Quantile plots. It compares your distribution to a stand distribution across quantiles and plot it.

Here we used it to see if the residuals when compared to normal distribution, does it hold up or not. If they match we should get almost all the points in the plot lying on the guide line.

https://www.analyticsvidhya.com/blog/2021/09/q-q-plot-ensure-your-ml-model-is-based-on-the-right-distributions/