Assignment-based Subjective Questions - Arun Gambhir

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

Following variable have impact on the dependent variable-

- yr
- holiday
- temp
- windspeed
- spring
- winter
- Jul
- Sep
- Mist
- Snow

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

Dummy variables create as many number of fields as there are types, but the combination of n-1 variable (consider n options) can tell the state of last variable so it will just add redundancy if we keep it. Hence removing it will help make the model more accurate with less variables.

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

Based on my graph which has temp, atemp, humidity & windspeed. Temp & atemp seems to be highly correlated with cnt (target variable)

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

Following are the activities I performed to validate the assumptions-

- Modify the model to remove the variable having p value greater than 0.05.
- Modify the model to remove the variable having VIF greater than 5.
- Did the residual analysis to get the mean at 0
- Also did the model evaluation to see a linear patter between actual & predicted value

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

Its temp, Snow & Year based on the coef value.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2531	0.024	10.569	0.000	0.206	0.300
yr	0.2342	0.008	28.210	0.000	0.218	0.251
holiday	-0.0980	0.026	-3.727	0.000	-0.150	-0.046
temp	0.4498	0.031	14.686	0.000	0.390	0.510
windspeed	-0.1395	0.025	-5.540	0.000	-0.189	-0.090
spring	-0.1123	0.015	-7.360	0.000	-0.142	-0.082
winter	0.0449	0.012	3.602	0.000	0.020	0.069
Jul	-0.0729	0.018	-4.167	0.000	-0.107	-0.039
Sep	0.0573	0.016	3.606	0.000	0.026	0.089
Mist	-0.0796	0.009	-9.014	0.000	-0.097	-0.062
Snow	-0.2855	0.025	-11.445	0.000	-0.334	-0.236

General Subjective Questions

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

Linear regression is a process to find the target variable from the independent variables by providing a coefficient to them. The algorithm is to find that correct value of these coefficients, so it gives the correct output.

Following are the steps to be followed –

1. We required the historical data based on which we can analyze & make the model learn.
2. Before talking about the model data needs to be cleansed & formatted so the model can learn from it.
    a. It requires- converting categorical values into dummy variables.
    b. Scaling the data so it gets converted in 0-1 range or mean value.
    c. Fixing the null values, duplicate data.
3. Then do EDA (Exploratory data Analysis) to understand the data how it varies with target values & among another field. In case a few fields are highly correlated we can drop one.
4. Now it's time to work on the model. The first step is to split data into test & train.

    a. On the train data we do fit & transform so model will learn & transform the data accordingly.
    b. Once done using statsmodel we add a constant to the data.
    c. Then apply the OLS to get the summary of model.
    d. If the p-value is greater than 0.05 or VIF value greater than 5, we repeat the above step else we take that as final model
5. Next step is to do the residual analysis to check graph is getting created around mean 0 or not.
6. Once done we can follow the same procedure for test data. The only difference is instead of Fit we follow transform here & use the model that we have created in training.
7. Using that model, we can evaluate test predicted & actual values like in a linear graph.

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

Anscombe quartet states that its always recommended to check the scatter plot of data before applying linear regression on them. Sometimes the data is not meant for linear regression.

To prove it Anscombe took 4 dataset whose Mean, variance, Standard Deviation and R square are all same but when they plot the data on scatter plot, they observed only one follows the linear format rest all were matching due to outliers or random combinations.

Conclusion- Visualization of data is the key & first step for regression analysis. So always follow that & make the data looks clean & then perform the analysis.

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

Pearson's R is used to define the relationship between two variables. Whether they are strongly/weakly/positively/negatively correlated with each other.

Following values are used to present these relationships-

r = 1, Strong Positive, r = -1, Strongly Negative

r = 0, No Correlation

r > .5, Positive, r < -0.5 Negative

0 < r < .3, Weak Positive, 0 > r > -0.3 Weak Negative

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

Scaling is done to make the value in similar range. So, coefficients can be applied to generate the prediction.

Support we have one field for 'number of rooms' in house & other 'area' of apartment. The value for a room could be 1,2,3 & area its 1500, 2000. So, to get coefficient for this is very hard when the difference is huge, so we scale the data so they both come in similar range like 0 to 1 or -1 to +1 etc...

There are 2 types of scaling-

Normalized = (value – min)/ max - min

Standard = (value – mean)/ standard deviation

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

VIF provides a measure of multicollinearity between the independent variables. This will adversely affect the regression result. The less the VIF, the less the dependency of one variable on the other.

When the VIF is infinite that means some variables are highly corelated to each other so it impacts the regression result hence we have to remove them one by one based on the VIF values.

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

Quantile-Quantile plots are used to find whether 2 sets have similar distribution or not. Suppose we got the train & test set differently.

So, to make sure their distribution aspects like location, scale & outliers are similar we use this plot.

This Q-Q plot holds the x & y quantiles, if both exist on 45-degree line then they have similar distribution. If they are far away from this line, then it's not.