**Review of Group Proposal by Linda Huang, Manav Kakar, Omer Tahir, and Yuchen Zhang**

This Review is broken up into a short summary and longer, detailed component. Overall, this was a *well designed project that was very clearly described*.

**Shorter Review:**

**Positives:**

* All parts of Reasoning are explicitly stated and it is very clear to reader why you do certain parts and what the purpose is. Methods section is very well written.
* Figures have excellent captions which explains the role of the picture and what the reader should focus on
* Great looking graphs and figures
* Code works fine (I was able to run it fine in RStudio but not JuptyerHub. It might be worth double checking in case its not just an error by me.)
* References well formatted and consistent.

**Possible Improvements:**

* Some minor parts of the introduction and Results sections were awkward to read and could be polished. I provided some examples and suggestions below.
* Figures and Tables should be numbered.
* Some more writing that explicitly tells us why the two research questions are interesting and why they are related enough to go into the same research project will make the goals more clear.
* More explicit information of why the research literature you read is relevant and what are the important findings from them that lead to the need for your research. This is important as right now it only feels like you are mentioning other research as a homework requirement and only at a surface level.
* Perhaps the first research question could be rephrased as it feels like a big claim to test:

“Are temperature, wind speed and precipitation ***the only weather conditions*** that significantly affect Bike Sharing Demand?” as the dataset you have might not have all the possible variables that could affect it. Maybe you could use a softer phrase?

Again, I think you did a great job and figures look great. These are just minor suggestions I feel would take it to the next level. Good work!

**Longer Review:**

**Mechanics:**

The provided ipynb file did not work via my JupyterHub HOWEVER the same code did work on RStudio and includes all libraries needed. Therefore, it appears the mechanics should work great.

**Reasoning:**

I think the reasoning for the project is clearly stated. You want to build a generative model and also a predictive model for bike sharing usage using LM.

I also really like how you explain explicitly what you wish to gain from the generative model (test which factors are most important to bike demand) and what the purpose of the second model is (prediction).

I also really like how you acknowledge how you will tackle problems to LM models.

**Writing:**

There are a few small improvements that could be made with the writing.

1. An interesting title to go with your interesting research
2. In the Introduction you wrote:
   1. “The usage of Bike Sharing can help reduce the consumption of natural sources and air pollution in sustainable cities”

I think this sentence can be made smoother. Did you mean natural resources (gasoline in particular)? You could write “… help reduce the consumption of natural resources and reduce air pollution to create sustainable cities” (as Bike Sharing could be started in cities that are not sustainable)

* 1. “we want to study the factors that”

This could be “we will study” to make more direct and confident

* 1. “One existing research from a university student group explored “the impact of economic, environmental and social factors for students cycling” to drive a well-thought design for the bike-sharing service [2]. Another research which we thoroughly scanned was focused more on predicting bike sharing demand using neural networks.”

I like how you found two studies to compare. However, I feel you did not go into enough meaningful detail about those studies and why their results or findings are useful for your project. Did you mean to say that these 4 factors are the most important ones you need to consider? If so you will need to explicitly say so.   
  
The quotation you provide is not the strongest one because it leaves the reader wanting to find out more about why this is important and those words are not unique enough that you could not have paraphrased it yourself. I feel quotes are best used when the authors have a special wording or a special point that is best described exactly the way they did.

To improve this, you might add on what these two studies found that is interesting in relation to your research or how it lead to your research questions in more detail (so its easy for reader to connect the steps your mind took).

“to drive a well-thought design” is an awkward statement. You may try “a student group explored [X, Y, Z] as important factors in a well thought out bike sharing design”

Or: “a student group found these factors [X, Y, Z] were important factors to a well thought out bike share design”

* 1. I feel improving this literature component will also make it easier for you to explain why these two research questions you chose are related and belong in the same report.
  2. Perhaps the first research question could be rephrased as it feels like a big claim to test:

“Are temperature, wind speed and precipitation the only weather conditions that significantly affect Bike Sharing Demand?” as the dataset you have might not have all the possible variables that could affect it. Maybe you could use a softer phrase?

1. For the Results section:
   1. Graphs and summaries should be labelled with a number, eg. Figure 1. Or Table 1. Followed by a descriptive title. I think you did a good job providing a descriptive title already.
   2. Excellent captions below your figures, I really liked how you spelled out the logic of what you were presenting:

Eg. “In general, an absolute correlation coefficient of >0.7 among two or more predictors indicates the presence of multicollinearity.

This suggests that the variables dew\_point\_temp and temp are indicating the presence of multicollinearity since they have a very high correlation of 0.91.”

1. For Methods Section:
   1. “we will check its goodness of fit using the training set”

Did you mean to right you will check it using the test set?

* 1. I am not sure you can conclude “temperature, wind speed and precipitation the only weather conditions that significantly affect Bike Sharing Demand” with the dataset you have as there are other weather variables not included.

Perhaps you could use a more precise and narrow description of what the generative model will produce.