**COGS 108 Previous Project Review (Wi21) - Notes**

The questions provided below are the questions included on the Google Form. Feel free to make a copy of this document and share with group members to collaborate on your answers. *However,* **you must copy + paste your answers to the Google Form for the actual submission**. **Each group will submit one Google Form.**

1. What is your group number? (i.e. 010) This can be found on your group's assigned GitHub repository.

030

1. What is the URL of the prior project you were assigned to review?

https://github.com/COGS108/FinalProjects-Sp19/blob/master/FinalProject\_Group102.ipynb

1. 32 word summary: In 32 words (or less), summarize the project.

Group 102's goal is to examine the differences between the different states of Kickstarter projects and some general patterns in companies that use Kickstarter. Ultimately, the group's analysis didn't support their expectations.

1. Considering the entire project (question, code, analysis, visualization, communication, etc.), explain the \*strengths\* of this project.

* The group’s data includes 378,661 projects, which is most likely good for making inferences.
* There is a ton of code provided in this project, making this analysis easily reproducible.
* In the Markdown cells, detailed explanations are given for almost every few lines of code, making their thought process clear and understandable.
* Entries with NaN values are singled out and scrutinized. This group provided sufficient explanation for why they believe all these entries are wrong entries, namely by comparing these entries to other valid entries and seeing distinct patterns in entries with those NaN values.
* The columns that are used in their analysis are quantified during their clean up process.
* The group caught two important things in this data set: outliers (decided to focus on the interquartile range) and a distribution of the goal values isn’t normal; however, there is one small issue (in the weaknesses).
* In the goal section, the group used histograms to visualize the data in all relevant columns, which is an excellent choice as it elucidates the different behaviors mentioned in the group’s descriptions for each column.
* The 5-number summary is provided for all the relevant columns, and for each column, this group points out all peculiar things in their histograms.
* This should be a given, but all parts of the question have been answered. In fact, the strongest part of the project is probably the Data Analysis and Results section because this is a very detailed and clear section. Some parts of the question have even been answered using multiple different approaches (for example, the analysis based on the goal of the Kickstarter projects was done in two different ways).
* The conclusion is well written. It’s succinct and condenses a very packed project that ultimately describes the information necessary to answer their questions (in this case, see that the data doesn’t match their expectations of the differences for the states of these projects)
* Limitation on information about the project owner is identified in the conclusion. The limitation is explained and possible extensions are presented.
* Although the data did not match the original expectation, the group proceeded on the research and investigated unexpected patterns.

1. Considering the entire project (question, code, analysis, visualization, communication, etc.), explain the \*weaknesses\* of this project.

* (Small weakness) The general background feels very chunky and long. It can probably be improved.
* (small weakness) The relevance between prior work and this project can be more explicitly described. For example, such as what can be learned from prior work that might provide useful information to this project.
* (Small weakness) This group should use more Markdown formatting to emphasize key terms such as “successful,” “failed,” etc.
* (Small weakness) In Data Cleaning, they shouldn’t be using the code block Markdown format for the list because the text wrapping feature of code block causes a lot of confusion.
* (Small weakness) This group did notice a discrepancy between 'usd pledged' and 'pledged' as well as 'goal' and 'usd\_goal\_real' and decided to use one of these types of categories (USD or the country’s currency type); however, when they pointed this out, they didn’t specify which type of category they used in their analysis (edit: it wasn’t until further down that they mentioned which type of column they used, which is USD. They should have mentioned this earlier for clarity).
* The group shows that almost all the Kickstarter projects are located in the US. The question could probably be narrowed down to focus only on US projects (and thus prune the entries that are not in the US) rather than projects in all countries. The group did account for differences in these different countries, but it probably would have been a lot better to just analyze the US because there isn’t that much data to use for the other countries. The good thing is that they did consider this issue a few times in the project, mainly in the Ethics and Privacy section.
* (Small weakness) In the graphic for visualizing the proportions of successful vs. unsuccessful campaigns for each country, they put the split bar for successful Kickstarter campaigns somewhere in the middle of each bar instead or either the top or bottom, which makes it slightly harder to compare the proportions of successful vs. unsuccessful campaigns for each country.
* The group only checked the distribution of the goal column and not the pledged column. It might be enough to just see that one of these distributions isn’t normally distributed, but I think it might still be better to examine the distributions of both of these columns.
* (Small weakness) The bar chart in the Success Rate Per Goal is a bit confusing because they used different colors (including the ones used in the legends of previous visuals), making me think something is missing. All bars in the chart should be blue for clarity. The rest of the visuals in the Data Analysis and Results section seem fine, though.
* When the group compares the percentage of each state of a particular project category over time, they only examine one category, which is technology. Although question 2 mentions that most of the Kickstarter projects are technology projects, the research would be more compelling if more than one category was examined.
* The Ethics and Privacy section looks unpolished and incomplete. Most of the description here is about bias rather than ethics and privacy themselves, and there seems to be a comment suggesting that more should have been added (“Gabe add here”).

1. Given what you've learned from this project, what aspects of this project do you want to \*keep/incorporate\* into your own work this quarter?

* Most of the group projects (not just the one we’re reviewing) don’t have a table of contents, but for a project as long as the one we’re reviewing, it would be easier to navigate if there is a table of contents near the top of the Github page.
* This group did very well to describe in detail their analysis for each of the subsections. I think we should do the same for our project, but not make it as long as group 102 did.
* A lot of the project outside of the Data Analysis and Results section is dedicated to cleaning the data, which we see is an important part of the project because it will reveal things that may help narrow down our question.
* Some of the types of graphs that seem particularly effective in visualizing data (especially multiple categories of data) are histograms, bar charts (including split bar charts), and line plots, which may serve to be a useful asset in our group’s project.
* Similarly to group 102, we should make it clear in the conclusion the information we obtained from parts of the project that specifically answer the question and accept or reject our hypotheses.

1. Given what you've learned from this project, what aspects of this project do you want to \*change/try to avoid\* into your own work this quarter?

* The questions are quite narrow. It seems to be difficult to narrow down the question in the early stages of the project, so what I think our group should do is narrow the question more as we go through the process of cleaning and analyzing data.
* Although I listed this as a strength, the group added information about the 5 number summary in places that didn’t seem to need it for the purpose of answering their question. They also did the same for figuring out if some categories of data are normally distributed. Unless this is required for the project, I don’t think we should add this to our project if they aren’t needed.
* Because ethics and privacy is a significant aspect of this research, it may be best to give this section more attention than what group 102 did and perhaps put this section near the top of the project in order to make it clear in the beginning.
* This may not be a big deal in the grand scheme of things, but perhaps we should use better Markdown formatting in our project.

1. If you were grading this project, what score (out of a possible 100 points) would you have awarded this project?

95

1. Explain briefly (1-3 sentences) why you would have awarded that score

Without a rubric, it’s difficult to quantify the work of this group, but given that this project is based on a class that assumes little no prior knowledge of data science, I think this is a decent project. When determining the score of this project (95), my decision is mainly based on the weaknesses, which most are negligible because they are small mistakes that are typical in a first-time project and sometimes don’t affect the content of the project. The point deductions mainly come from the unpolished ethics and privacy section (a topic that Professor Ellis greatly emphasizes, although this might not have been heavily emphasized in previous iterations of this class, so I’ll give some leniency), although I did take off a few points for some small issues in the data analysis.