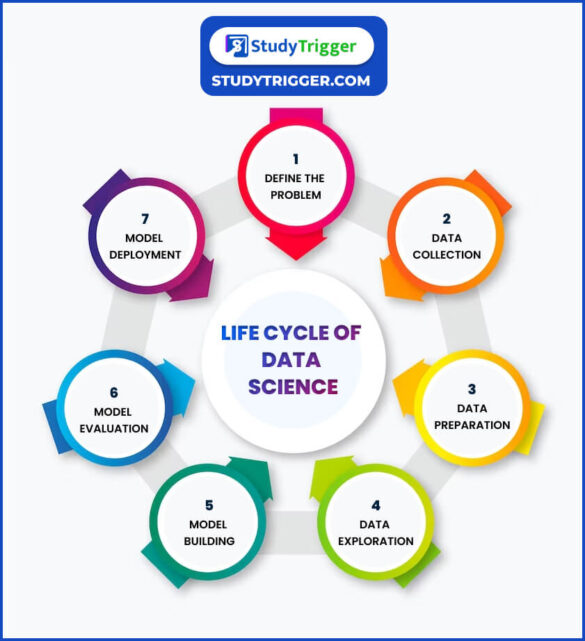
****

The group portion of the final project will include a group project [**proposal**](#_mytkdgoo3g1c), a group **write-up** in a Jupyter notebook (with visualizations and functioning code) and a group **poster** presentation.

Your final project group **write-up** must:

1. Discuss (but not necessarily implement) the 7 stages of the data science lifecycle as they relate to your topic of exploration. Stages 4, 5, and 6 will comprise the bulk of your project, but all the sections should be addressed. For example:
   1. If data is already collected, the data collection section could discuss concerns or important aspects of the methodology that should have been used when collecting the data.
   2. The data preparation section might describe the steps taken to ensure the data is in a useful form.
   3. The model deployment section might propose how the model could be deployed and/or potential concerns if it actually was deployed.
2. Investigate a hypothesis or a question.
3. Build and evaluate at least one model.
   1. The model doesn’t have to be predictive (for example, linear regression or decision tree that we cover in the last unit). A model just has to make some reasonable assumption on the data, and it doesn’t have to be complicated.
   2. For evaluation, you can use a technique appropriate for the model you have chosen, such as simulation, sampling, comparing distributions, or calculating accuracy. You might also evaluate the effects of outliers.
   3. You are also welcome to build multiple models and compare how they perform.
4. Investigate at least 3 variables and their potential associations.
5. Write at least one Jupyter Notebook file with at least 2 custom functions.
6. Create at least 3 original Python visualizations with appropriately labeled axes and titles.

Your final project group **poster** will be presented to the class during our scheduled finals time slot. It should include the most important elements from your write-up and should use visualizations effectively to “tell the story” of your data and your findings. All team members must participate in presenting the poster and should be prepared to explain any portion, including:

1. The problem being investigated.
2. The visualizations, why they were chosen, and how they were generated.
3. The model, why it was chosen, and how it was evaluated.

All of these datasets should appear in your **F24-public** folder when you **Git -> Pull from Remote**.

**🔥**: Expected level of difficulty **solely** based on the size and how “processed” the dataset is.

[**Track A: Penguins 🔥 3**](#_9lfqyvykrkrc)

[**Track B: Global issues 🔥 4**](#_1f5yucvgsz6k)

[**Track C1: Health and medicine - Diabetes 🔥🔥 5**](#_2set6ptx7xxw)

[**Track C2: Health and medicine - Algorithmic risk scores 🔥🔥🔥 7**](#_2p573cdoj16h)

[**Track D: Climate 🔥🔥 8**](#_sy07wurp2d2o)

[**Track E: Sports 🔥🔥🔥 9**](#_juto7tspmbvv)

[**Track F1: Business - Marketing campaign 🔥 11**](#_ojswfasykzpk)

[**Track F2: Business - Price optimization 🔥🔥 12**](#_9316ry76j70q)

[**Track G: Government 🔥🔥🔥 13**](#_m91qchi2onfk)

[**Track H: Suggest your own 🔥🔥🔥 14**](#_2tszmf99g7tq)

[**Project proposal (80 points, 8% of final course grade) 15**](#_mytkdgoo3g1c)

## 

## Track A: Penguins 🔥

If you are new to programming and statistics, we recommend this track! You would be able to build on Lab 10.

Our good old friend, penguins! This dataset was collected and made available by [Dr. Kristen Gorman](https://www.uaf.edu/cfos/people/faculty/detail/kristen-gorman.php) and the [Palmer Station, Antarctica LTER](https://pallter.marine.rutgers.edu), a member of the [Long Term Ecological Research Network](https://lternet.edu/). It was originally published in:

Gorman KB, Williams TD, Fraser WR (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE 9(3):e90081. <https://doi.org/10.1371/journal.pone.0090081>

**Sample questions:** What are the defining characteristics of each species? If you met a new penguin, how would you know which species it is? Which variables are the most important? When is this classification easy or difficult? What associations do you observe?

|  |  |
| --- | --- |
| **Variable** | **Value** |
| studyName | Sampling expedition from which data were collected, generated, etc. |
| Sample Number | the continuous numbering sequence for each sample |
| Species | the penguin species |
| Region | the region of Palmer LTER sampling grid |
| Island | the island near Palmer Station where samples were collected |
| Stage | reproductive stage at sampling |
| Individual ID | unique ID for each individual in dataset |
| Clutch Completion | if the study nest observed with a full clutch, i.e., 2 eggs |
| Date Egg | a date denoting the date study nest observed with 1 egg (sampled) |
| Culmen Length (mm) | the length of the dorsal ridge of a bird's bill (millimeters) |
| Culmen Depth (mm) | the depth of the dorsal ridge of a bird's bill (millimeters) |
| Flipper Length (mm) | an integer denoting the length penguin flipper (millimeters) |
| Body Mass (g) | an integer denoting the penguin body mass (grams) |
| Sex | sex of an animal |
| Delta 15 N (o/oo) | the ratio of stable isotopes 15N:14N |
| Delta 13 C (o/oo) | the ratio of stable isotopes 13C:12C |
| Comments | additional relevant information for data |

## Track B: Global issues 🔥

Explore **Gapminder** (<https://www.gapminder.org/data>) and **Our World in Data** (<https://ourworldindata.org/data>). Most of these datasets are time series data, or data for different countries, or both. If there are any datasets that you find interesting, download them as csv files. This process should be straightforward by clicking on buttons on the website, but please let us know on Piazza if you have any trouble.

This project comes with a starter Jupyter Notebook where I show you how to make geographic visualizations based on country names and ISO codes. You are ***not*** required to make map visualizations to receive full credit for this project. This is just in case you are interested.

**Sample questions:** Do you see any associations between variables that are common across many countries? Do you see any global trends in time for any variables? Are there any countries that are extreme outliers from observed associations? For example, exploring happiness: is a country’s happiness related to its population? or its median age? Is global happiness changing over time? What other variables appear correlated to happiness?

## Track C1: Health and medicine - Diabetes 🔥🔥

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

diabetes.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. This dataset has 21 feature variables.

**Sample questions:** What are the defining characteristics of people with diabetes? Or with prediabetes? If you are a physician presented with a new patient, how would you diagnose them based on their data? Which variables are the most important in making that diagnosis? What kind of advice or intervention would you give to folks at risk of diabetes?

More details and other people’s code working with this dataset:

<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

Original dataset from CDC: <https://www.cdc.gov/brfss/annual_data/annual_2015.html>

|  |  |
| --- | --- |
| **Variable** | **Value** |
| Diabetes\_012 | 0 = no diabetes, 1 = prediabetes, 2 = diabetes |
| HighBP | 0 = no high blood pressure, 1 = high BP |
| HighChol | 0 = no high cholesterol, 1 = high cholesterol |
| CholCheck | 0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years |
| BMI | Body Mass Index |
| Smoker | Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes |
| Stroke | (Ever told) you had a stroke. 0 = no 1 = yes |
| HeartDiseaseorAttack | coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes |
| PhysActivity | physical activity in past 30 days - not including job 0 = no 1 = yes |
| FruitsConsume | Fruit 1 or more times per day 0 = no 1 = yes |
| Veggies | Consume Vegetables 1 or more times per day 0 = no 1 = yes |
| HvyAlcoholConsump | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes |
| AnyHealthcare | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes |
| NoDocbcCost | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes |
| GenHlth | Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor |
| MentHlth | Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days |
| PhysHlth | Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days |
| DiffWalk | Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes |
| Sex | 0 = female 1 = male |
| Age | 13-level age category (\_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older |
| Education | Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate) |
| Income | Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more |

## Track C2: Health and medicine - Algorithmic risk scores 🔥🔥🔥

This data is from the following article:

Ziad Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366,447-453(2019). <https://doi.org/10.1126/science.aax2342>

In this article, the authors use patient medical records, demographics, and insurance claims to study bias in a machine learning model used to predict patient risk. This model has been used to make recommendations about which patients should be admitted to more intensive care programs on the basis of their health. The results presented in this article are discussed by Dr. Ruha Benjamin in the video "[Are We Automating Racism](https://www.youtube.com/watch?v=Ok5sKLXqynQ)" around minute 14, which could be a good place to start.

**Sample questions:** Which variables contribute the most to determining a patient’s algorithmic risk score? Can you reproduce the qualitative findings of this paper? Note that this is just an example. Your project does not have to be related to the research questions presented in the original paper.

Data: <https://gitlab.com/labsysmed/dissecting-bias/-/raw/master/data/data_new.csv?inline=false>

**Variables** (it was too long to copy): <https://gitlab.com/labsysmed/dissecting-bias/-/blob/master/data/data_dictionary.md?ref_type=heads>

## Track D: Climate 🔥🔥

The Global Historical Climatology Network monthly (GHCNm) dataset provides monthly climate summaries from thousands of weather stations around the world. The initial version was developed in the early 1990s, and subsequent iterations were released in 1997, 2011, and most recently in 2018. The period of record for each summary varies by station, with the earliest observations dating to the 18th century. Some station records are purely historical and are no longer updated, but many others are still operational and provide short time delay updates that are useful for climate monitoring. The current version (GHCNm v4) consists of mean monthly temperature data, as well as a beta release of monthly precipitation data. NCEI uses GHCN monthly to monitor long-term trends in temperature and precipitation.

Original dataset (the servers may not work because of the hurricane 🙁) : <https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-monthly>

This project comes with a starter Jupyter Notebook (plotly.ipynb) where I show you how to make geographic visualizations based on latitude and longitude data. You are ***not*** required to make map visualizations to receive full credit for this project. This is just in case you are interested. If you run the code in climate.ipynb, you can download data from more decades starting 1901.

**Sample questions:** Do you observe any long term (or short term) trends in the temperatures? Any differences based on geographical locations of the stations? How does the average yearly change in temperature vary within a given country? **IF** everyone in your team is interested and feeling confident, you can try to combine storm and hurricane data from <https://catalog.data.gov/dataset/ncdc-storm-events-database2> or climate data from [Track B](#_1f5yucvgsz6k).

|  |  |
| --- | --- |
| **Variable** | **Value** |
| ID | Station identification code. First two characters are [FIPS country code](https://en.wikipedia.org/wiki/List_of_FIPS_country_codes) |
| NAME | Station name |
| LATITUDE | Latitude of station in decimal degrees |
| LONGITUDE | Longitude of station in decimal degrees |
| STELEV | The station elevation in meters. -999.0 = missing. |
| YEAR | 4 digit year of the station record. |
|  | monthly value (MISSING=-9999). Temperature values are in hundredths of a degree Celsius, but are expressed as whole integers (e.g. divide by 100.0 to get whole degrees Celsius). |

## Track E: Sports 🔥🔥🔥

Since 2019, NFL has been hosting the [NFL Big Data Bowl](https://operations.nfl.com/gameday/analytics/big-data-bowl/):

The annual sports analytics contest from NFL Football Operations challenges members of the analytics community – from college students to professionals – to contribute to the NFL’s continuing evolution of the use of advanced analytics. The crowd-sourced competition uses data and technology to spur innovation that results in creating new insights, making the game more exciting for fans and protecting players from unnecessary risk.

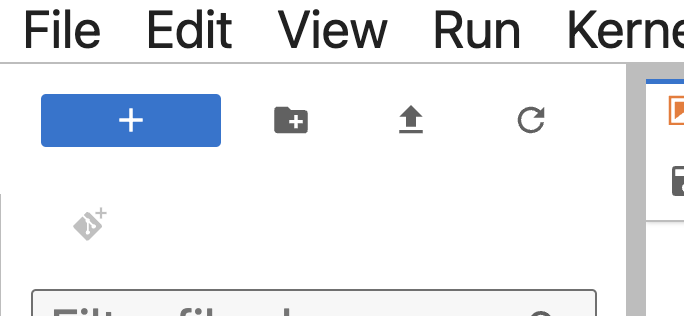
We are going to work with a smaller, beginner-friendly version of the 2019 Big Data Bowl data, downloaded from: <https://www.kaggle.com/datasets/aryashah2k/beginners-sports-analytics-nfl-dataset>

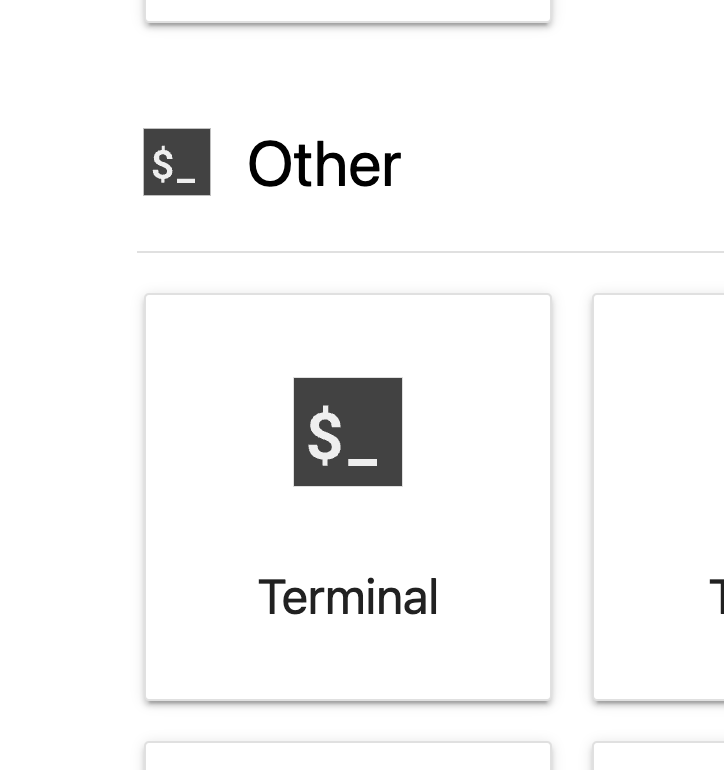
It has four tables, each saved in a different file. You can see descriptions of each variable on the kaggle website. This project comes with a starter Jupyter Notebook file where I show you how to “join” or “merge” tables based on unique IDs. You also saw an example in an earlier homework.

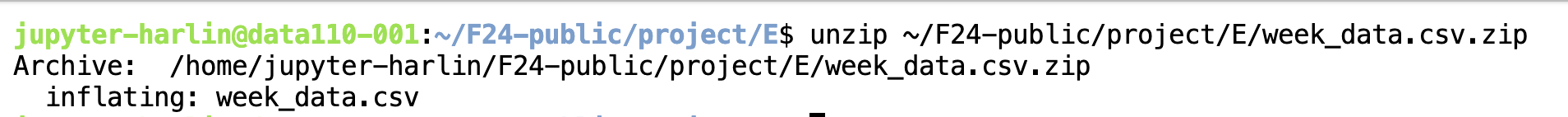
**Sample questions:** Do you observe any relationship for running certain plays, or certain types of plays, depending where on the field you are? or whether you are home or away? or depending on the physical attributes of your roster? or what point in the season it is? Do you observe any variable greatly affecting the yardage result of plays? What variables affect the play of players the most? How would you quantify these things?

**IF** everyone in your team has a lot of experience from previous classes and would prefer to work on a different sport analytics dataset, you can try to work on the non-beginner friendly version of the dataset, or you can browse this website: <https://sportsandsociety.osu.edu/sports-data-sets>. Please make sure to run it by us first before writing your proposal.

See next page on how to unzip week\_data.csv.zip:

1. Press the blue button in top left corner. ****
2. Click on Terminal button.

****

1. Type the following “unzip ~/F24-public/project/E/week\_data.csv.zip” and hit enter. ****

## 

## Track F1: Business - Marketing campaign 🔥

[iFood](https://en.wikipedia.org/wiki/IFood) is the lead food delivery app in Brazil, present in over a thousand cities. Keeping a high customer engagement is key for growing and consolidating the company’s position as the market leader. Data Analysts working within the data team are constantly challenged to provide insights and value to the company through open scope projects. This case intends to simulate that. In this case, you are presented a sample dataset, that mocks metainformation on the customer and on iFood campaign interactions with that customer.

See “iFood Data Analyst Case.pdf” for context and sample questions, but remember that you don’t need to answer all (or any) of their questions. Data downloaded from: <https://github.com/nailson/ifood-data-business-analyst-test>. This person also uploaded their approach to the problem, which you are welcome to study for inspiration.

|  |  |
| --- | --- |
| **Variable** | **Value** |
| AcceptedCmp[1/2/3/4/5] | 1 if customer accepted the offer in the [1/2/3/4/5]th campaign, 0 otherwise |
| Response (target) | 1 if customer accepted the offer in the last campaign, 0 otherwise |
| Complain | 1 if customer complained in the last 2 years |
| DtCustomer | Date of customer’s enrollment with the company |
| Education | Customer’s level of education |
| Marital | Customer’s marital status |
| Kidhome | Number of small children in customer’s household |
| Teenhome | Number of teenagers in customer’s household |
| Income | Customer’s yearly household income |
| Mnt[FishProducts/MeatProducts/Fruits/SweetProducts/Wines/GoldProds] | Amount spent on [fish/meat/fruits/sweet/wines/gold] products in the last 2 years |
| NumDealsPurchases | Number of purchases made with discount |
| NumCatalogPurchases | Number of purchases made using catalog |
| NumStorePurchases | Number of purchases made directly in stores |
| NumWebPurchases | Number of purchases made through company’s website |
| NumWebVisitsMonth | Number of visits to company’s website in the last month |
| Recency | Number of days since the last purchase |

## Track F2: Business - Price optimization 🔥🔥

A pricing experiment has been done by company ABC with the main aim of increasing revenue. In the experiment, the users were grouped into two groups A (66% of the user base) and B (33% of user base), where in group A users were offered a cheaper price while group B users received a higher price.

**Sample questions:**

* What are your main findings looking at the data? What is your overall view into user behavior, especially focusing on actionable insights that might increase conversion rate?
* What is your recommendation to the company in terms of setting an optimum price for their product?
* Can you optimize the number of days that the test is run? After how many days would you have stopped the test? Why?

This project comes with a starter Jupyter Notebook file where I show you how to “join” or “merge” tables based on unique IDs. You also saw an example in an earlier homework.

This was an interview question for a data scientist role at Amazon Sydney. Data and prompts downloaded from: <https://github.com/nailson/amazon_ds_challenge>. **IF** everyone in your team has a lot of experience from previous classes and would prefer to work on the more difficult challenges (2 and 3 in the repository), that is ok with us (in principle). Please run it by us first before writing the proposal.

## Track G: Government 🔥🔥🔥

For this track, you can pick your own dataset! We found three sources for you that include datasets about Chapel Hill or North Carolina.

* Chapel Hill datasets on data.gov: <https://catalog.data.gov/dataset/?q=&sort=views_recent+desc&groups=local&_organization_limit=0&organization=town-of-chapel-hill-north-carolina>
* NC office of state budget and management: <https://linc.osbm.nc.gov/explore/?sort=modified>
* Data sources curated by the NC office of state budget and management: <https://www.osbm.nc.gov/open-data-resources/open>

**Sample questions:** There is a huge variety of data here, so many different types of questions and exploration of associations are possible, so be creative! A couple examples: for Chapel Hill, is there a relationship between air quality and fire incidents? Do any variables seem to have an observable time cycle? Is there a measurable trend over time for any variables?

## Track H: Suggest your own 🔥🔥🔥

If you already have a project you’re working on for a different class or as a research assistant or for your portfolio, just ask us. We have to approve it before you start writing your proposal! In principle, we are okay with it as long as all members in your team are equally excited about it and have the necessary tools and skills to equally contribute to the project.

# Project proposal (80 points, 8% of final course grade)

Please make a group submission on Gradescope: <https://guides.gradescope.com/hc/en-us/articles/21863861823373-Adding-Group-Members-to-a-Submission>

Make sure to include a project title(1 point) and a (cool) team name (1 point)!

Expected length: 1~2 pages, 400~500 words. A little longer or shorter is okay.

1. Define the problem (20 points)
   1. What is a question you have about the world that this project could potentially help answer? Or what is a problem or unmet need that this project can help with? Write down the “big picture” question and why you are interested in it.
2. Data collection (10 points)
   1. Even if you are working with a dataset we have provided, try to describe the origin of this dataset in your words. Who collected the data? Who is the population? Comment on potential limitations and biases. If you wish, you are welcome to collect your own data through, for example: surveys, counting the number of people who visit Meantime Coffee, or measuring the arrival times of Chapel Hill transit. If you intend to collect your own data, include a detailed plan for doing so.
3. Data preparation (10 points)
   1. Even if you are working with a dataset we have provided, it is likely that you would have to prepare them in some manner for your specific project. Refer to lecture 11 slides for examples. If this step is not relevant, say so.
4. Data exploration (20 points)
   1. Given what we have learned in this class so far, you already have the tools to calculate summary statistics and visualize the data. Include some of those preliminary results (a few numbers and/or a figure is fine) and comment on what you have learned about the data. No need to share your code. Refer to lecture content for examples. Your data exploration may help add specificity to part 1.
5. Model building (10 points)
   1. You’re not expected to have done any work or have a very detailed plan for model building at this point. But given what you have learned about the data from step 4, do you have any assumptions on data that you would like to explore more?
6. Risks and timeline (8 points)
   1. What are potential risks or failures you foresee in this project? How would you proceed if that happens? What is the planned timeline for this project? Remember that all assignments related to this course have to be submitted by the final exam date.