**Contact Information:**

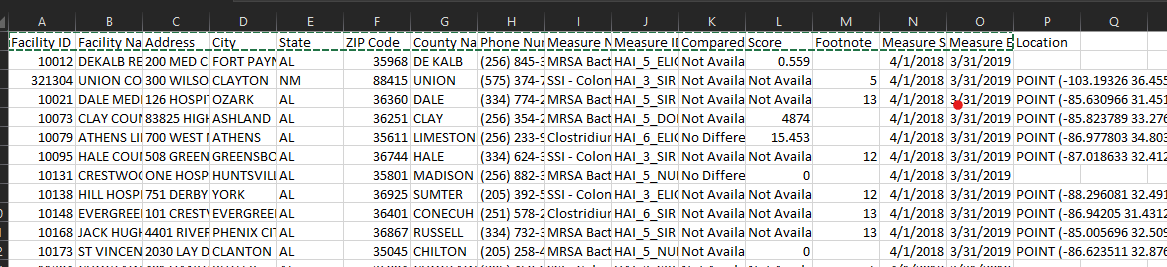
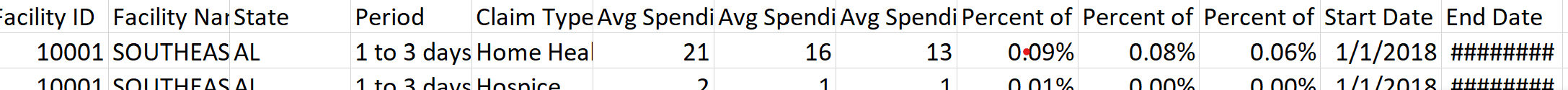
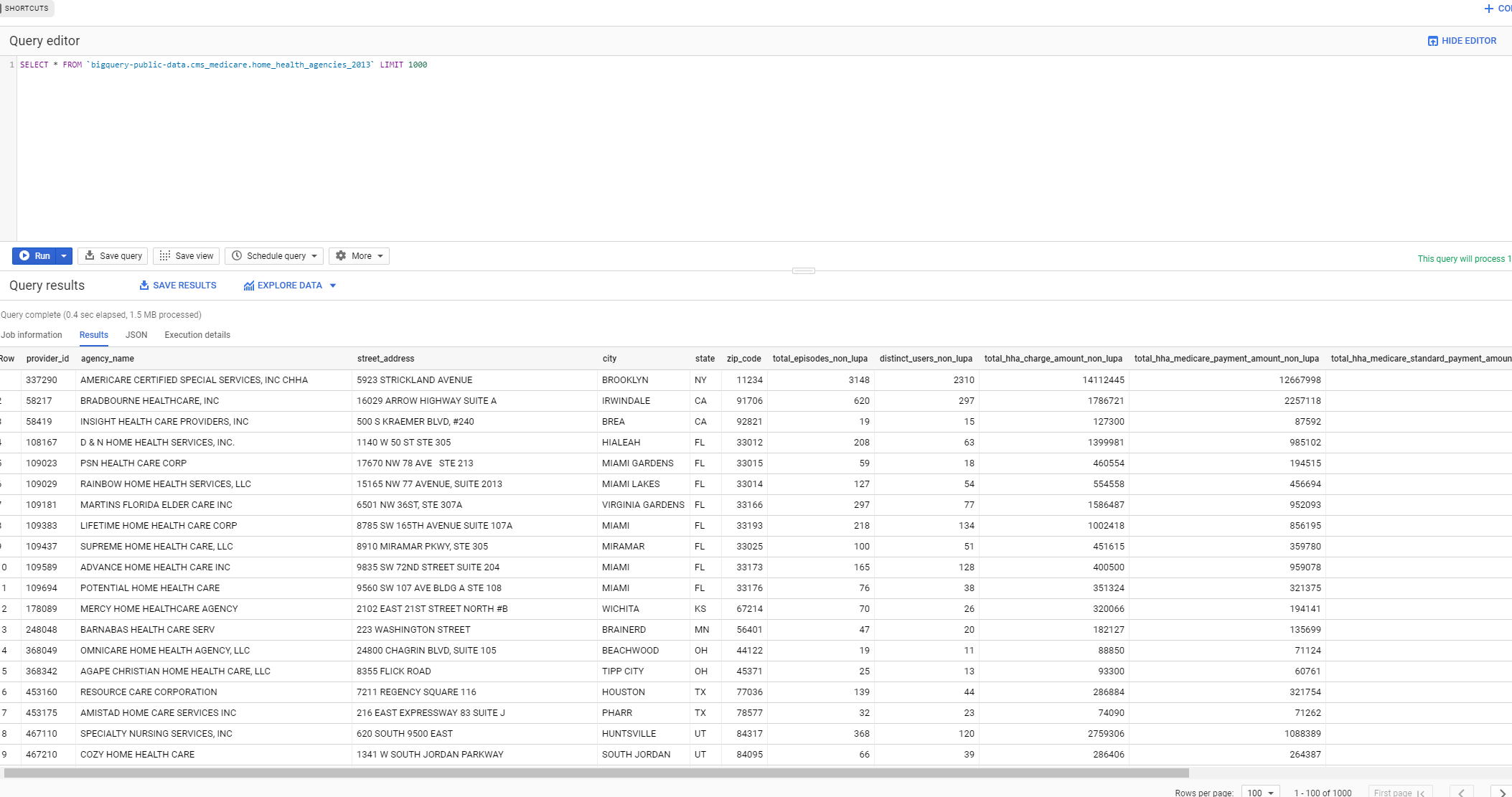
|  |  |  |
| --- | --- | --- |
| **Name** | **Email** | **Mobile** |
| Salik Hussaini | salikhussaini49@gmail.com | 224-713-5374 |
| Pete Gillespie | gillespie.peterj@gmail.com | 952-693-1929 |
| Zoya Hafeez | hafeez.zoya@gmail.com | 267-585-3029 |

**Guidelines:**

Final Project ProposalBefore you dive into building out your final project, I’d like you to craft a proposal outlining your data, problem definition and goals. This should allow you to clarify what you’re looking to build, as well as get feedback on that objective before committing your time it would take to build it.

By the time you submit your proposal you should have formed your team and have decided on a dataset with some practical goals other than just analysis. You should clearly state why you are interested in this data and what problems you will try to solve. The problem should be clearly defined. You should write at least four research questions or queries that you will investigate.

The purpose of the proposal is to help you channel your thinking in the right path by asking questions on your data:

1. **What? What is your problem, what is your data, what is your expertise?**
   1. **Problem Statement:**
      1. Healthcare is important as age increases. Disease burden exponentially takes a toll on an individual’s life. Visiting a healthcare provider also increases, so in result debt also increases. What we aim to do is reduce billing and unneeded ICD-10 codes applied to a patient’s visit. The benefits of our analysis is a reduction of a debt of a patient's hospital visit.
      2. United is notorious for high healthcare costs, many politicans preach that the healthcare system is broken. With these ideas in mind, our group will focus on solving why is healthcare so expensive? What is actually going wrong? Our data from the Us. Department of health and human services provides common inpatient and outpatient services, all physician and other supplier’s procedures and services, and also Part d prescription. With this immense data source we hope to apply several machine learning algorithms to combat the high healthcare costs.
      3. **Option 3:** 
         1. **Predictive analysis of healthcare cost to help reduce debt accumulation from medical providers.**
         2. [**https://data.medicare.gov/Hospital-Compare/Healthcare-Associated-Infections-Hospital/77hc-ibv8**](https://data.medicare.gov/Hospital-Compare/Healthcare-Associated-Infections-Hospital/77hc-ibv8)
         3. [**https://data.medicare.gov/Hospital-Compare/Medicare-Hospital-Spending-by-Claim/nrth-mfg3**](https://slack-redir.net/link?url=https%3A%2F%2Fdata.medicare.gov%2FHospital-Compare%2FMedicare-Hospital-Spending-by-Claim%2Fnrth-mfg3)
         4. [**https://console.cloud.google.com/marketplace/details/hhs/medicare?filter=solution-type%3Adataset&filter=category%3Ahealth&id=8daf3f67-990c-48ad-8ee4-bf3874794550**](https://slack-redir.net/link?url=https%3A%2F%2Fconsole.cloud.google.com%2Fmarketplace%2Fdetails%2Fhhs%2Fmedicare%3Ffilter%3Dsolution-type%253Adataset%26filter%3Dcategory%253Ahealth%26id%3D8daf3f67-990c-48ad-8ee4-bf3874794550)
      4. ****
      5. ****
      6. ****
2. **Why? Why is this important, what’s the benefit?**
   1. **Importance:**
      1. **For Option 1, Measuring Treatment Effectiveness**: This application of data mining involves comparing and contrasting symptoms, causes and courses of treatment to find the most effective course of action for a certain illness or condition. For example, patient groups who are treated with different drug regimens can be compared to determine which treatment plans work best and save the most money. Furthermore, the continued use of this application could help standardize a method of treatment for specific diseases, thus making the diagnosis and treatment process quicker and easier.
      2. **For Option 2, Detecting Fraud and Abuse**: This involves establishing normal patterns, then identifying unusual patterns of medical claims by clinics, physicians, labs, or others. This application can also be used to identify inappropriate referrals or prescriptions and insurance fraud and fraudulent medical claims. The Texas Medicaid Fraud and Abuse Detection System is a good example of a business using data mining to detect fraud.
      3. Option 3
         1. [Ref: <https://archer-soft.com/en/blog/data-mining-healthcare>]
      4. Healthcare is important as age increases. Disease burden exponentially takes a toll on an individual’s life. Visiting a healthcare provider also increases, so in result debt also increases. What we aim to do is reduce billing and unneeded ICD-10 codes applied to a patient’s visit. The benefits of our analysis is a reduction of a debt of a patient's hospital visit.
3. **How? How will you get the data, how will you solve the problem, what tools will you use?**
   1. > Give some background on the data: Once we finalise on our data set, we can probably include related fields in the proposal?
   2. > Provide a snapshot of the dataset: Prof wants a snapshot of data? Is he referring to data dictionary or just snapshot?
   3. Our data comes from one source is the cms.gov website, whichso
4. **How much? How deep will you go, how much analysis will you do?**
   1. I will help you with whether it is OK to continue or make some modifications to your proposal. The problem you want to work shouldn’t be too easy, but also shouldn’t be too hard. The dataset should be sufficiently large and complex. At the proposal stage you have done basic exploratory analysis on different datasets, tried to answer the above questions, and finally decided on a data set of your liking.

Going into a bit more detail, the proposal should be approximately one to three pages long and answer the following questions:

* What is the problem you are attempting to solve?
* How is your solution valuable?
* What is your data source and how will you access it?
  + **oGive some background on the data - Salik Hussaini**
  + **oProvide a snapshot of the dataset - Salik Hussaini**
* What techniques do you anticipate using, and why? --
  + numerical, categorical, geographic, demographic
    - Linear Regression, PCA, hypothesis testing.
* What do you anticipate being the biggest challenge you’ll face?
  + Having an R-squared that is low,
    - however is still able to adequately predict future medical services.
  + Multicollinearity
  + Assumption not met for regression model.
* What research question (queries) are you going to investigate?
  + If i were to go to a hospital which provider would bill me the least?
  + If i had a medical condition which city will bill me the least.

When answering these questions you should form a clear picture of the work you intend to do without having begun to build out the infrastructure to execute it yet.

You may have written some code, done some initial scraping, or some initial analysis, even some simple models. Do not, however, start to actually build your product until it has been approved.

Other than the above questions, I expect to see some basic exploratory analysis including some statistical summaries and data visualizations.

**Submit:Maximum three page long proposal as Word document**

* **What? What is your problem, what is your data, what is your expertise?**
  + United is notorious for high healthcare costs, many politicans preach that the healthcare system is broken. With these ideas in mind, our group will focus on solving why is healthcare so expensive? What is actually going wrong? Our data from the Us. Department of health and human services provides common inpatient and outpatient services, all physician and other supplier’s procedures and services, and also Part d prescription. With this immense data source we hope to apply several machine learning algorithms to combat the high healthcare costs.
* **Why? Why is this important, what’s the benefit?**
  + Healthcare is important as age increases. Disease burden exponentially takes a toll on an individual’s life. Visiting a healthcare provider also increases, so in result debt also increases. What we aim to do is reduce billing and unneeded ICD-10 codes applied to a patient’s visit. The benefits of our analysis is a reduction of a debt of a patient's hospital visit.
  + **I expect to see some basic exploratory analysis including some statistical summaries and data visualizations.**
* **How? How will you get the data, how will you solve the problem, what tools will you use?**
  + Our data comes from one source is the cms.gov website, whichso
* **How much? How deep will you go, how much analysis will you do?**
  + What techniques do you anticipate using, and why? --
    - numerical, categorical, geographic, demographic
      * Linear Regression, PCA, hypothesis testing.
  + What do you anticipate being the biggest challenge you’ll face?
    - Having an R-squared that is low,
      * however is still able to adequately predict future medical services.
    - Multicollinearity
    - Assumption not met for regression model.
  + What research question (queries) are you going to investigate?
    - If i were to go to a hospital which provider would bill me the least?
    - If i had a medical condition which city will bill me the least.