**PROJECT PROPOSAL**

**Team Name:** Datanauts  
**Team Members:** Glen Cancian, Siddharth Sahai, Sonal Verma, Vivek Pawar

**Motivation for the project:**

A health insurance provider is looking to break in to the United States market and it has hired Datanauts to undertake market analysis. Ideally, they would like to target between 1 & 3 counties with their health insurance products. The business questions identified below will assist them in understanding the opportunities available in the US healthcare market and enable them to best position their products & services. In addition, the insurer also wants to know which demographics (race and origin) of the society should be targeted in the selected states/counties.

Most insurance providers do not include an option for autopsy as part of their insurance schemes due to the high cost. In case of foul play, the laws differ from state to state but in many circumstances the relatives or next of kin or the person themselves, as part of their will, request an autopsy. Given our Client is looking to break into the health insurance market, they would like to offer coverage for autopsy and thus want to know if this is something that may give them an edge over other health insurance providers.

**Insights and learnings sought:**

Hypothesis:

1 – Over the past five years, persons over 65 years of age are more likely to have died whilst at home of a natural death than any other type of death.

Questions:

1 - Which States are those that have lower levels of health insurance coverage (gap in the market) and have higher mortality rates (county level).

2 - What are the three age brackets with the highest death rates over the past 5 years and what seasons were most prevalent?

Data Mining:

1- Are there clusters of States that are particularly susceptible to chronic diseases that relate to disease? (subsets of the chronic disease indicators file). Are there any clusters for race and Hispanic origin?

2- Is there an association between age and death by Parkinson's or Alzheimer's?

3- If there is a death reported then based on the parameters predict the classification of the person in terms of race and origin? (perform supervised learning using classification)

4- Based on other parameters predict if a death needs an autopsy? (perform supervised learning using classification on all columns to classify autopsy)

**Technical approaches:**

To test the proposed hypothesis, we will employ statistical analysis techniques in R, Python or Minitab. Classification and association algorithms will be used to discover the best insights for each of the data mining business questions. These will be implemented in either R or Python. Tableau and/or R will be used to visualise findings from descriptive analytics including the answers to the two Questions identified above.

**Data Explanation:**

Four individual data sets have been sourced from Kaggle.com. Each of the datasets is identified in the table below. Note that the first three datasets will able to be joined on US State, whilst the fourth dataset will be used as a standalone dataset for market trend analysis and insight.

|  |  |  |
| --- | --- | --- |
| **Name** | **Link** | **Record Count** |
| Mortality Rates | https://www.kaggle.com/IHME/us-countylevel-mortality | 67,075 |
| Health Insurance Coverage | https://www.kaggle.com/hhs/health-insurance | 53 |
| Chronic Disease Indicators | https://www.kaggle.com/cdc/chronic-disease | 403,984 |
| Death in the United States | https://www.kaggle.com/cdc/mortality | 2.5 - 2.7 mill / year |

**Discuss project execution risks:**

R1 – Incompatibility of Datasets: there may be some incompatibility of chosen datasets given that these are all separate datasets.

R2 – Quantum of Data: the combined amount of data may not be easily and efficiently processed.

R3 – Python: none of our team members are familiar with the appropriate algorithms written in Python. We are only familiar with those implemented in WEKA.

R4 – Learnings from this course may find an application in the project but it may be too late to implement them. Most of the methods speculated as of now are based on our knowledge from the Data Mining course.

**Risk mitigation strategies:**

R1 – Incompatibility of Datasets: a specific task to confirm compatibility will be assigned to a team member at initial commencement of our project.

R2 – Quantum of Data: Appropriate data from datasets 3 & 4 will be subset for use.

R3 – Python: once our project proposal is accepted, we will assign a team member to commence sourcing/writing appropriate algorithms in Python.

R4 – We will follow start-up methodology of *Learn to Fail Early*. This will give us flexibility in terms of finding out if our strategy is bound to fail before it is too late.

**Specify goals and deliverables:**

|  |  |
| --- | --- |
| **Goal** | **Deliverable** |
| Undertake market research and develop strategies for market entry | Documented insights to each of the business questions identified above. |
| Developing business understanding and defining success criteria. | Project Proposal |
| Data Exploration and Data Quality Check | Data Quality and Exploration Report |
| Data flattening | Flattened datasets and subset as required |
| Hypotheses Testing and Iterative Modelling | Statistical and Data Mining models |
| Evaluating project against success criteria | Documented evaluation |
| Creating visualization, report review and finalization | Final project submission including visual dashboard(s) and all findings |