Alex Kaplan

Project Outline

Task 6.1 Sourcing Open Data

**Summary and Source:** The world has had many citizens who have had high medical costs even with insurance. Some medical costs with insurance can be over $50,000. With all these high medical costs, it would be beneficial to look at data to see if trends or patterns exist that help us predict if a patient will have high medical costs or not. This insurance dataset shows demographic information regarding a patient such as their age, BMI, and how many children they have.

**Collection:** The insurance dataset was collected from Kaggle which contained demographic information on 1337 from people in four geographic regions in the United States.

**Contents**: Contains one dataset: Insurance Dataset

**Ethics:** The data abides by HIPAA laws and does not put anyone’s data at risk. Data bias does not seem to be an issue since we do not have much personal information from the patient that would impact our judgment.

**Limitations:** The insurance data was collected manually so there could be some errors. There are also only 1339 rows of data which may not be enough of a sample to predict insurance costs for every patient. We also do not know specifically what city or state a patient lives in. The data just provides regions of the US. Much of this data is limited to patients with access to healthcare, hospitals, and willingness to seek medical treatment so there could be even more insurance costs than this dataset suggests.

**Reason for Choosing Dataset:** I chose this dataset because I am very interested in how medical insurance costs are determined. I also thought it would be interesting to find common factors that influence a patient’s medical costs.

**Consistency Checks/Cleaning Data – No duplicates or missing values found in the Insurance Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Missing Values | Missing Value Treatment | Duplicates | Inaccuracies/Other |
| None | None | 1 | There was a duplicated row that was removed from the dataset |
|  |  |  | Charges was changed to ‘insurance\_charges’ to make it more consistent with the dataset. |

**Insurance Dataset**

|  |  |
| --- | --- |
| Column Name | Description |
| Age | Age of the patient |
| Sex | Sex of the patient male or female |
| bmi | Body mass index of the patient measuring weight to height. |
| Children | Number of children covered by insurance/number of dependents |
| Smoker | Whether a patient smokes or not. yes or no |
| Region | The beneficiary’s residential area in the US: northwest, southeast, southwest, |
| Charges | Individual medical costs billed by health insurance |

**Data Profile**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Time Variant/Invariant | Structured/Unstructured | Quantitative/Qualitative | Nominal/Ordinal/Discrete/Continuous |
| Age | Invariant | Structured | Quantitative | Discrete |
| Sex | Invariant | Structured | Qualitative | Nominal |
| BMI | Time Variant | Structured | Quantitative | Continuous |
| Children | Time Variant | Structured | Quantitative | Discrete |
| Smoker | Invariant | Structured | Qualitative | Nominal |
| Region | Invariant | Structured | Qualitative | Ordinal |
| Charges | Variant | Structured | Quantitative | Continuous |

**Questions to Explore:**

1. Which regions have the highest insurance charges?
2. Is there a connection between age and insurance charges?
3. Is there are connection between insurance charges and someone’s BMI?
4. What impact does smoking have on someone’s insurance charges?

[GitHub Repository](https://github.com/akap0209/Insurance-Charges-Analysis)