1. ABSTRACT & DATASET

Summary of Final Project for "344SM - Statistical Methods"

- **Topic**: Analysis of the Australian health survey dataset (1977-1978).
- **Data**: Dataset comprises 5190 individuals and 19 variables, focusing on 'doctorco' (number of consultations with a doctor or specialist in the past 2 weeks).
- **Objectives**:
    1. Initial data exploration, including data cleaning and visualization.
    2. Predicting the number of doctor/specialist consultations in the past 2 weeks.
- **Methodologies Used**:
    - Statistical Models: Negative Binomial Regression, Zero-inflated Poisson Regression, Zero-inflated Negative Binomial Regression, Hurdle Poisson Regression, and Hurdle Negative Binomial Regression.
    - Machine Learning Models: Logistic Regression, Random Forest, Naive Bayes, Neural Networks, and K-Nearest Neighbors.
- **Programming Languages**:
    - R: Utilized for its advanced statistical analysis and visualization capabilities.
    - Python: Employed for superior data manipulation, machine learning, and integration features.
- **Project Approach**:
    - The project utilized a dual-language strategy, combining R and Python to leverage the strengths of both.
    - This approach enabled efficient handling of various stages of the project, from data preprocessing to complex modeling and deployment.
- **Benefits**:
    - Enhanced quality and depth of analysis.
    - Flexible and innovative approach for addressing complex project challenges.
    - Maximization of the capabilities of both programming ecosystems for a comprehensive analytical workflow.

Summary of Variables in Health Survey Dataset

- **Categorical (Binary) Variables**:
    - 'sex': Indicates gender (1 for female, 0 for male).
    - 'levyplus': Indicates private health insurance coverage (1 for covered, 0 for not).
    - 'freepoor': Indicates government coverage due to socio-economic status (1 for covered, 0 for not).
    - 'freepera': Coverage due to old-age, disability, or veteran status (1 for covered, 0 for not).
    - 'chcond1': Presence of chronic condition without activity limitation (1 for yes, 0 for no).
    - 'chcond2': Presence of chronic condition with activity limitation (1 for yes, 0 for no).

- **Categorical (Ordinal) Variables**:
    - 'age': Age represented as a fraction of 100.
    - 'agesq': Square of the 'age' variable.
    - 'income': Annual income in thousands of dollars, coded into ranges.
    - 'hscore': Health score from a general questionnaire, higher scores indicating worse health.
- **Discrete Variables**:
    - 'illness': Count of illnesses in the past 2 weeks, with a cap at 5.
    - 'actdays': Days of limited activity due to illness or injury in the past 2 weeks.
    - 'doctorco': Number of doctor or specialist consultations in the past 2 weeks (response variable).
    - 'nondocco': Consultations with non-doctor health professionals in the past 2 weeks.
    - 'hospadmi': Hospital admissions in the past 12 months, capped at 5.
    - 'hospdays': Nights spent in a hospital during the most recent admission, with a maximum coded value of 80.
    - 'medicine': Total medications used in the past 2 weeks.
    - 'prescrib': Prescribed medications used in the past 2 weeks.
    - 'nonpresc': Nonprescribed medications used in the past 2 weeks.

These variables collectively help to profile health service utilization patterns, health status, and socio-economic dimensions of the individuals in the survey, offering a rich dataset for analysis.

2. Considerations

Summary: Healthcare in Australia during 1977-1978 and the Medibank Scheme

- **Context**: The Australian healthcare system in 1977-1978 was undergoing significant changes, primarily due to the implementation of Medibank, a universal health insurance system.
- **Medibank Introduction**:
    - Introduced in 1975 by the Whitlam Government through the Health Insurance Act 1973.
    - Aimed to provide universal health coverage to all Australians and free treatment in public hospitals.
    - Funded by a 2.5% levy on taxable incomes, an additional levy for high-income earners, and government funds.
- **Transition in Healthcare System**:
    - In 1981, the Fraser Government abolished Medibank, replacing it with Medibank Private, a government-subsidized private insurance scheme.
    - The dataset analyzed in this project corresponds to the period when the original Medibank scheme was operational.
- **Healthcare Structure**:
    - Australia's healthcare system is federally structured, dividing responsibilities between federal and state/territory governments.
    - The federal government handled funding and policy aspects, while state/territory governments managed healthcare service delivery.

- **Challenges and Debates**:
    - The cost of running Medibank was a significant concern for the government.
    - There were debates on whether Medibank achieved equitable access for all Australians.
    - Concerns were raised about longer waiting times in the public healthcare system.

This period represents a pivotal moment in Australian healthcare history, with Medibank being a crucial element in shaping the future direction of the nation's health policy and insurance systems.

3. RESPONSE VAR, IND VAR, CONSIDERATIONS

Summary of Exploratory Data Analysis (EDA) Approach

- **Dataset Integrity**: Despite its age, the dataset is complete with no missing values and is considered clean.
- **Initial Exploration**:
    - The distribution of the response variable 'doctorco' will be examined.
    - The distributions of other variables will also be analyzed.
- **Correlation Analysis**:
    - Investigation into the correlation between the variables and the response variable 'doctorco'.
- **Data Simplification**:
    - The variable 'agesq' (age squared) is excluded from the analysis.
    - This is done to simplify the models, enhance interpretability, and avoid problems like redundancy and multicollinearity.
    - The exclusion is not expected to significantly affect the accuracy of the analysis.

The EDA will focus on understanding the variable distributions and their relationships, particularly how they relate to the frequency of doctor consultations.

Summary: Analysis of the Response Variable 'Doctorco' in a Healthcare Dataset

- **Variable Description**: 'Doctorco' represents the number of doctor or specialist consultations in the past 2 weeks.
- **Data Distribution**:
    - The distribution is highly skewed to the right.
    - A significant majority (79.8%) reported zero consultations, indicating a low rate of doctor consultations in this timeframe.
    - A smaller proportion of individuals reported one or more consultations.
- **Data Visualization**:
    - **Histogram**: Shows a high frequency for 0 consultations, dominating other counts. This reflects the skewness towards fewer doctor visits.
    - **Density Plot/Rug Plot**: Likely used to show individual data points, highlighting the concentration of data at 0 and sparser distribution for higher consultation numbers.
    - **Log-Transformed Histogram**: Adjusts the scale to better represent the distribution of non-zero consultation counts, making the long-tail of higher consultation numbers more visible.
- **Implications for Healthcare**:

- The distribution and mean values of 'doctorco' are vital for healthcare planning and resource allocation.
- Insights from this data help understand the utilization patterns of medical services.

This analysis provides a clear picture of the consultation behavior in the dataset, emphasizing the predominance of no consultations within a two-week period and the rarity of higher consultation counts. Such information is crucial for effective healthcare management and policy-making.

Summary of Health Survey Variables

- **Gender (‘sex’)**: Slight female majority with approximately 52% of the sample.
- **Age (‘age’)**: Range from 19 to 72 years old; median age of 32 suggesting a distribution slightly skewed towards older individuals.
- **Income (‘income’)**: Data divided by 1000 and coded into ranges, with the most common income level being 0.25.
- **Private Health Insurance (‘levyplus’)**: 2,298 individuals (about 44.2% of the sample) have private health insurance coverage for a private patient in a public hospital.
- **Government Health Care (‘freepoor’)**: 222 individuals (approximately 4.2%) covered by government healthcare due to low income, recent immigration, or unemployment.
- **Old-age/Disability Pension (‘freepera’)**: 1,091 individuals (approximately 21.02%) covered for health care due to old-age or disability pension.
- **Illnesses (‘illness’)**: On average, individuals reported about one to two illnesses in the past two weeks.
- **Reduced Activity Days (‘actdays’)**: Average of 0.86 days of reduced activity in the past two weeks, with 85.8% reporting no days of reduced activity.
- **Health Score (‘hscore’)**: Mean score is 1.218, suggesting generally good health with 58.3% reporting no health issues.
- **Chronic Conditions without Limitation (‘chcond1’)**: About 40.31% of individuals have chronic conditions without limiting their activities.
- **Chronic Conditions with Limitation (‘chcond2’)**: 605 individuals (about 11.66%) have chronic conditions that limit their activities.
- **Non-doctor Consultations (‘nondocco’)**: Very few consultations with non-doctor health professionals; heavily skewed towards zero consultations.
- **Hospital Admissions (‘hospadmi’)**: 86.5% reported no hospital admissions in the past year, with 10.8% having one admission.
- **Hospital Days (‘hospdays’)**: 86.5% had no nights in a hospital; smaller proportions reported 1 to 80 nights with higher frequencies for shorter stays.
- **Medication Use (‘medicine’)**: 42.9% did not use any medication in the past two days; on average, individuals used just over one medication.
- **Prescribed Medications (‘prescrib’)**: 59.4% did not use any prescribed medications in the past two days; average use is less than one prescribed medication.

- **Non-prescribed Medications (‘nonpresc’)**: 73.5% did not use any non-prescribed medications; on average, one-third of a medication was used in the past two days.

These variables collectively provide a comprehensive overview of the health status and healthcare utilization among the survey participants, indicating a prevalence of good health with low utilization of healthcare services, but with a significant minority having chronic conditions and healthcare needs.

Summary: Insights and Considerations from Health Data Analysis

- **Key Insights**:
    - The dataset provides a detailed perspective on healthcare utilization.
    - Binary variables such as gender, insurance status, and chronic condition presence allow for straightforward modeling and clear population segmentation.
    - Count variables like 'illness' and 'actdays' are informative for modeling healthcare demand.
    - The response variable 'doctorco' shows most individuals did not have doctor consultations, which aids in planning and policy-making.
- **Challenges**:
    - The right-skewed 'doctorco' distribution suggests potential zero-inflation issues for statistical models.
    - Discretized variables like age and income may result in information loss versus continuous data.
    - Long-tailed distributions in 'hospdays' and 'medicine' may necessitate specialized models to address overdispersion.
    - Variables with limited variance might not significantly enhance model complexity or accuracy.
- **Modelling Considerations**:
    - Careful consideration is needed in the selection and modeling of variables to ensure a quality analysis.
    - The real-world dataset exemplifies common data challenges, emphasizing the importance of comprehensive analysis in understanding context.
- **Potential for Broader Analysis**:
    - Inclusion of variables related to other health services, emergency service usage, and preventive health services could yield a more complete view of health service utilization.
    - Certain health conditions, such as stroke or diabetes, could be explored for their impact on healthcare service use.
- **Frequency of Health Services**:
    - There's a potential consideration that not all health services are required within a 2-week window; some may be needed less frequently.
- **Tools Used for Analysis**:
    - The use of both R and Python showcases the flexibility and robustness of the analysis, reflecting the expertise of the analysts.

This analysis underscores the complex nature of healthcare data and the necessity of methodical approaches to uncover meaningful insights that can inform health service delivery and policy.

4. BIVARIATE ANALYSIS

CORRELATION MATRIX

Summary of Correlation Matrix Interpretation

- Key Correlations Identified:
    - **Healthcare Utilization**: There is a positive association between age and health issues with the frequency of doctor consultations, indicating greater healthcare needs in these groups.
    - **Income and Healthcare**: Lower-income individuals tend to have fewer doctor visits, suggesting possible access barriers for this demographic, even when health issues are present.
    - **Complex Health Management**: There is a trend where older individuals with more health issues have longer hospital stays, indicative of the complexity in treating multiple or severe health conditions.
    - **Medication Usage**: A higher number of health issues correlates with an increased prescription of medications, reflecting the necessity of medical management for these conditions.
    - **Non-Prescribed Medication Use**: An increase in the use of non-prescribed medications is noted among older individuals with more health problems and doctor visits.
- **Considerations**:
    - **Correlation and Causation**: The observed correlations do not imply causation, and there may be other underlying factors affecting these patterns.
    - **Healthcare Disparities**: The disparities related to income and healthcare access highlight a need for further research into the socioeconomic factors affecting healthcare utilization.

The findings provide a groundwork for understanding healthcare patterns and indicate areas where policy interventions may be required, particularly in addressing disparities in healthcare access.

BIVARIATE ANALYSIS WITH DOCTORCO

Summary of Healthcare Utilization Factors

- **Age**: Older individuals have more doctor consultations, reflecting greater healthcare needs as age increases.
- **Sex and Income**: Certain demographic groups, differentiated by sex and income levels, show varying frequencies of doctor visits, hinting at disparities in health conditions, access to care, or health-seeking behaviors.
- **Healthcare Coverage**:
    - **Private Insurance ('levyplus')**: No significant correlation with doctor consultations, indicating that private insurance status does not strongly influence medical visit frequency.
    - **Government Coverage ('freepoor', 'freepera')**: Significant associations, particularly with 'freepera', suggest that government-supported

individuals, such as older adults, those with disabilities, or veterans, are more likely to have more doctor consultations.

- **Health Episodes ('illness', 'actdays'):** A clear link exists between the number of illnesses and activity-limiting days with the frequency of medical visits, underlining that health problems significantly drive healthcare seeking behavior.
- **Perceived Health ('hscore')**: Poorer health scores correlate with more frequent doctor consultations, aligning with the likelihood of seeking medical attention when feeling unwell.
- **Non-doctor Professional Interactions ('nondocco')**: Increased interactions with non-doctor health professionals are related to more doctor consultations, possibly reflecting comprehensive care needs or more complex health conditions.
- **Chronic Conditions ('chcond1', 'chcond2')**: A strong association, especially with 'chcond2', suggests that chronic conditions, particularly those that limit activity, lead to more doctor visits.
- **Hospital Utilization ('hospadmi', 'hospdays')**: Both the number and length of hospital stays are associated with a higher frequency of doctor consultations, indicating that hospital events are typically followed by increased outpatient care.
- **Medication Use ('medicine', 'prescrib', 'nonpresc')**: Prescribed medication use correlates strongly with more doctor consultations, implying the necessity of ongoing medical supervision. However, the use of non-prescribed medication is not significantly associated with doctor visit frequency, suggesting that self-medication practices might not result in higher healthcare utilization.

This analysis paints a comprehensive picture of the determinants of doctor consultation frequency, encompassing demographic factors, health status, and healthcare engagement patterns.

5. Variable combination - Interaction effect on response variable

Quick Summary of Variable Effects on Doctor Consultations

1. **Activity Days and Illness**: More 'actdays' and illnesses lead to more doctor visits. Interaction is highly significant ($p < 0.001$).
2. **Age and Medications**: Older age and more prescriptions increase consultation frequency. Interaction greatly improves model fit.
3. **Income and Government Coverage**: Low income with government healthcare correlates with more doctor visits. Interaction is significant ($p = 0.0080$).
4. **Income and Private Insurance**: Higher income and private insurance impact consultation frequency. Significant model improvement ($p = 0.0050$).
5. **Sex and Health Score**: The effect of health score on doctor visits differs by sex. Interaction effect is significant ($p = 0.0149$), requiring further analysis for detailed interpretation.

These results underscore the importance of considering how different variable combinations influence the frequency of doctor consultations.

6. First cluster analysis

Cluster Analysis and Elbow Method
Cluster Analysis in Healthcare:

- **Patient Grouping: Identifies similar patient groups for better disease and treatment insights.**
- **Disease Subtyping: Aids in tailoring treatments, especially in complex diseases like cancer.**
- **Resource Management: Helps allocate resources effectively, prioritizing high-risk patients.**
- **Predictive Analytics: Uncovers patterns for forecasting health trends.**
- **Healthcare Delivery: Informs the design of targeted healthcare programs.**
- **Risk Stratification: Segregates patients into risk categories for proactive healthcare.**
- **Medical Insights: Discovers new data correlations, useful in genomics and epidemiology.**
- **Cost Efficiency: Reduces healthcare costs, especially in chronic disease management.**
- **Performance Benchmarking: Allows comparison and improvement of healthcare outcomes.**

Elbow Method in Cluster Analysis:

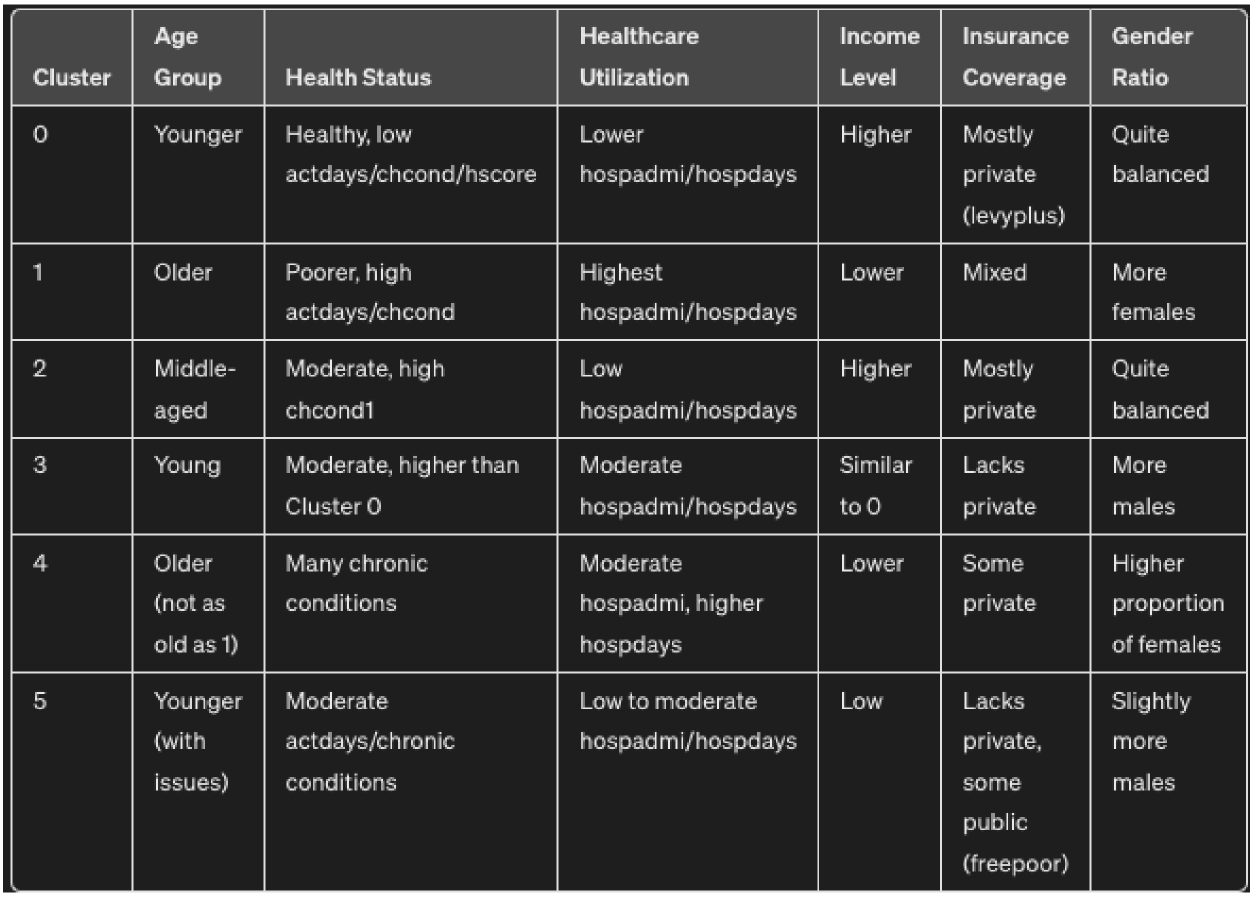
- **Optimal Clusters: Determines the best number of clusters for data analysis.**
- **K-Means Clustering: Uses K-Means with a range of cluster values and identifies the 'elbow' point as the optimal k.**
- **Python Implementation: Employs feature selection, standardization, and KElbowVisualizer for finding optimal k.**
- **Elbow Point: At k=6, the rate of decrease in distortion score changes, indicating the optimal cluster count.**
- **Centroids and Clustering: Re-runs K-Means with optimal k and assigns labels to data, using centroids for cluster profiling.**

Cluster analysis is essential for segmenting healthcare data and the Elbow Method is key for finding the ideal number of clusters to use in such analyses.

CLUSTER PROFILING

Maybe show a table in presentation

To profile the clusters, it was examined the centroid values for each clusters. These values gives an idea of the "typical" member of each cluster.



Cluster	Age Group	Health Status	Healthcare Utilization	Income Level	Insurance Coverage	Gender Ratio
0	Younger	Healthy, low actdays/chcond/hscore	Lower hospadmi/hospdays	Higher	Mostly private (levyplus)	Quite balanced
1	Older	Poorer, high actdays/chcond	Highest hospadmi/hospdays	Lower	Mixed	More females
2	Middle-aged	Moderate, high chcond1	Low hospadmi/hospdays	Higher	Mostly private	Quite balanced
3	Young	Moderate, higher than Cluster 0	Moderate hospadmi/hospdays	Similar to 0	Lacks private	More males
4	Older (not as old as 1)	Many chronic conditions	Moderate hospadmi, higher hospdays	Lower	Some private	Higher proportion of females
5	Younger (with issues)	Moderate actdays/chronic conditions	Low to moderate hospadmi/hospdays	Low	Lacks private, some public (freepoor)	Slightly more males

Each of these profiles suggests different needs and characteristics. Strategic decisions can be made based on these insights. For example, preventive health measures may be prioritized for clusters with chronic conditions but not currently utilizing a lot of healthcare services (like Cluster 2). On the other hand, policy makers may focus on providing better economic support or healthcare access to clusters like Cluster 5, who might be economically disadvantaged.