**Executive Summary**

This project involved the development and implementation of a K-means clustering application using a patient dataset to segment patients based on healthcare coverage and expenses. The dataset had 3540 rows and several columns including gender, ethnicity, birthplace, address, etc. The goal was to identify unique patients’ segments to improve healthcare delivery and resource allocation.

**Overview of the Code**

The application was developed using Python in the PyCharm IDE. The steps involved in the code development include the following:

1. Data preprocessing
2. Feature engineering
3. K-means clustering
4. Visualization
5. Unit testing

**Challenges and solutions**

1. Handling date columns: the original dataset had the birthdates in a date format. They were converted to ages and made numeric for easily clustering (Singh, 2019).
2. Deciding optimal number of clusters: K-means application was developed to group patients into unique clusters based on their healthcare expenses and coverage. The *Elbow Method* was used to determine the number of clusters (Carrascosa, 2024).
3. Categorical data encoding: columns containing categorical data such as marital status, race, ethnicity and gender could not be analysed until they were encoded.
4. Unit testing: extensive unit testing was conducted to ensure the application could handle invalid data and potential errors (Maatalla, 2023).

**Interesting Findings**

1. Unique patient segments: the K-means clustering successfully grouped patients into unique segments based on the specified criteria. The segments can be helpful with optimizing resources allocation and improving planning.
2. Improved insights: the cluster provided valuable information about patients’ healthcare coverage and expenses, helping providers to provide personalized healthcare plans and increase patients’ satisfaction.

**Conclusion**

This project demonstrated the effectiveness of K-means clustering in healthcare analysis and provided meaningful insights to improve healthcare services and patient care.

References

Carrascosa, I. P. (2024, September 27). *Performing cluster analysis in python: A step-by-step tutorial*. Statology. https://www.statology.org/performing-cluster-analysis-in-python-a-step-by-step-tutorial/

Maatalla, A. (2023, November 20). *Effective error handling: Preventing and handling exceptions*. Medium. https://medium.com/@abedmaatalla/effective-error-handling-preventing-and-handling-exceptions-06100aa36373

Singh, H. (2019, December 26). *Data Normalization for Numeric features - Analytics Vidhya - Medium*. Analytics Vidhya. https://medium.com/analytics-vidhya/data-transformation-for-numeric-features-fb16757382c0