Vaccine Hesitancy Project: Code Evaluation Document

---

1. Use of Python and Analytical Libraries

Overview: The project effectively utilised Python and its analytical ecosystem to analyse vaccine hesitancy data.

Key Libraries Used:

- **Pandas**: Used for data manipulation, cleaning, and pre-processing.
- **NumPy**: Applied for numerical calculations and handling missing data.
- **Matplotlib**: Leveraged for bar charts, scatter plots, and heatmaps.
- **Seaborn**: Created visually appealing plots such as correlation heatmaps.
- **Plotly**: Produced advanced interactive visualizations like Sankey diagrams and circular barplots.

Code Evaluation:

- Functions and libraries were employed efficiently, demonstrating proficiency in data science workflows.
- Visualizations were created with clear labelling, legends, and colour schemes for interpretability.

---

2. Defined Stages

2.1. Data Sourcing:

- Data was sourced from a **credible organization**: the UK **Office for National Statistics (ONS)**.
- Dataset: Coronavirus and Vaccine Hesitancy, January–July 2021.
- This dataset is reliable and representative of the UK population.

2.2. Pre-Processing:

- Cleaned and standardized the dataset to address:
    - Missing values.
    - Inconsistent categorical labels.
    - Formatting issues (e.g., column headers, data types).
- Calculated new metrics such as:
    - Hesitancy to Sentiment_Proportion.
    - Statistical significance using a chi-square test.

2.3. Evaluation:

- Performed robust analysis including:
    - Identifying predictors of vaccine hesitancy and positive sentiment.
    - Comparing categorical and numerical features.
    - Using statistical significance tests to evaluate relationships.

2.4. Visualization:

- Presented findings through a variety of visualizations:
    - Bar charts to highlight top predictors.
    - Heatmaps to display correlations.
    - Circular barplots and dumbbell plots for intuitive comparison.
    - Sankey diagrams for hierarchical data flow representation.

2.5. Data Transformation

Creating a New DataFrame:

- A new DataFrame was created to consolidate and analyse key metrics derived from the original dataset.
- **Purpose**:
    - To focus on predictors with statistically significant relationships to vaccine hesitancy or positive sentiment.
    - To create a simplified structure for specific visualizations like Sankey diagrams or dumbbell plots.

Steps Taken:

1. Filtered the original dataset for significant predictors (e.g., predictors with p-value < 0.05).
2. Added derived metrics:
    - Hesitancy to_Sentiment_Proportion: Ratio of vaccine hesitancy to positive sentiment.
    - -log10(p-value): Log-transformed significance values for easier interpretation.
3. Organized predictors into hierarchical categories for visualization purposes (e.g., "Income" → £20-39k, £40-59k).

Output:

- The new DataFrame allowed for:
    - A focused analysis of significant predictors.
    - Streamlined input to visualizations like Sankey diagrams and bar plots.

---

3. Relevance of Analysis Topic

Core Objectives:

- Determine key predictors of vaccine hesitancy and positive sentiment.
- Explore the relationship between demographic factors and attitudes towards vaccination.

Research Questions:

1. What are the strongest predictors of vaccine hesitancy?
2. How do predictors of positive sentiment differ from hesitancy predictors?
3. How does the distribution of sentiment and hesitancy vary across demographic groups?

Achievement:

- Objectives were clearly defined, and analysis provided actionable insights, such as the significant roles of income, mental health, and homeownership.

---

4. Data Sources

Primary Source:

- ONS: Coronavirus and Vaccine Hesitancy dataset.

Data Expansion:

- Explored augmentation opportunities such as incorporating additional survey data for mental health and regional information but kept focus on original dataset.

Documentation:

- All data processing steps and transformations were documented.

---

5. Visualization

Evaluation:

- Produced meaningful visualizations with clear annotations, legends, and titles.
- Examples:
    - Heatmaps demonstrated correlations between sentiment, hesitancy, and proportions.
    - Sankey diagrams illustrated the flow between predictors and outcomes.
    - Dumbbell plots compared predictors across sentiment and hesitancy.

Strengths:

- Visualizations were tailored to highlight specific insights.
- Plots were adjusted for accessibility (e.g., colourblind-friendly palettes, clear fonts).

---

6. Handling Data Anomalies

Approach:

- **Missing Values**: Handled by imputing where reasonable or excluding rows for categorical inconsistencies.
- **Outliers**: Examined and validated their relevance. For example, extreme hesitancy in certain subgroups was not treated as an anomaly.
- **Formatting Errors**: Adjusted column headers and ensured all numeric columns were standardized for consistency.

---

7. Credibility of Analysis

Findings Linked to Evidence:

- Statistical analysis, such as chi-square tests, supported conclusions.
- Visualizations consistently reflected the numerical findings, such as the relationship between income and hesitancy.

Example:

- The heatmap showed a strong positive correlation between vaccine hesitancy and lower income groups, validated through statistical significance testing.

Transparency:

- Code and methodology were documented, ensuring reproducibility.

Next Steps / Future Directions

1. Refinements:

- Incorporate more detailed geographic or time-based breakdowns.
- Explore causal relationships through logistic regression or machine learning.

2. Communication:

- Present results in a public-facing dashboard or interactive notebook for stakeholders.

3. Ethical Considerations:

- Include analysis of bias in the dataset (e.g., over-representation of certain demographics).