Exploratory Data Analysis Using Python on Health, Food, and Media Datasets

This analysis uses three real-world datasets — **Starbucks Menu**, **PIMA Indian Diabetes**, and **Netflix Titles** — to explore statistical distributions, relationships, and trends using Python. Key Python libraries used include `pandas`, `matplotlib`, `seaborn`, `plotly.express`, and `geopandas`.

1. Data Loading and Preparation

The datasets were loaded using `pandas.read_csv()` and explored using `.head()`, `.info()`, and `.describe()`. Columns were cleaned where needed (e.g., trimming whitespace in `starbucks`).

2. Univariate Analysis

- A **histogram** and **boxplot** for Starbucks `Calories` revealed a right-skewed distribution with outliers.
- Netflix content types were visualized using `countplot`, highlighting that **Movies** dominate the platform.
- The **age distribution** in the PIMA dataset was analyzed using KDE and histogram overlays, showing a concentration around age 30–50.

3. Bivariate Analysis

- A **scatter plot** of `Calories vs Sugars` in Starbucks showed a strong positive trend.

- Boxplots for `Glucose` across diabetes outcomes in PIMA indicated higher glucose levels among diabetic individuals.

- Netflix titles were analyzed over time, showing trends in movie and TV show releases by year using line plots.

---

4. Multivariate Analysis

- `BMI vs Age` was visualized using `scatterplot`, with color (`Outcome`) and size (`Pregnancies`) used for deeper insight into diabetes patterns.

- A **correlation heatmap** of PIMA features (e.g., `BMI`, `Glucose`, `Pregnancies`, etc.) helped identify relationships, such as a moderate correlation between `Glucose` and `Age`.

---

5. Subplot Grid

A 2×2 subplot layout integrated:

- Blood pressure histogram,

- Calories vs Protein scatter plot,

- Beverage category count plot,

- Age vs Outcome boxplot,
  with an overall title summarizing the multiview analysis.

---

6. Bonus Interactive Visuals

Using **Plotly Express**, interactive versions of the Starbucks boxplot and scatter plot were created, allowing zooming, tooltips, and filtering. This greatly enhanced exploratory flexibility.

---

7. Geospatial Mapping

Using **GeoPandas**, a district-level map of **Andhra Pradesh** was plotted from a `.geojson` file, showcasing the learner's ability to handle geographic data and shape visualizations.

Q: What does the histogram and boxplot tell us about Starbucks `Calories`?

- The histogram shows that most items have between **100–400 calories**, with a right-skew.
- The boxplot confirms **outliers** on the higher calorie end (above 500+).

Q: What is the most common type of content on Netflix?

- From the countplot, **Movies** are clearly more frequent than TV Shows in this dataset.

Q: What does the KDE and histogram tell us about Age in the PIMA dataset?

- Most patients are between **20 to 50 years old**.
- There's a relatively smooth, bell-shaped curve with a tail beyond age 60.

Q: What is the relationship between `Calories` and `Sugars (g)`?

- The scatter plot shows a **positive correlation** — higher calorie items tend to contain more sugar.

Q: What does the glucose boxplot suggest?

- Diabetic individuals (Outcome = 1) tend to have **higher glucose levels** than non-diabetics.

Q: How did Netflix trends change over the years?

- The number of titles increased steadily, peaking around 2018–2020.
- Movies were consistently released more than TV Shows.

Q: What did you learn from the BMI vs Age plots?

- Diabetic patients (Outcome = 1) often have **higher BMI**.
- The bubble size plot shows that **higher pregnancies** are clustered in older age groups.

Q: From the correlation heatmap, which features are most/least correlated?

- `Age` and `Glucose`: moderately correlated.
- `Pregnancies` and `BloodPressure`: weak or no correlation.

Q: What was useful in Plotly's interactive features?

- Hover data made it easy to inspect individual values.
- Zoom and pan features helped focus on dense areas.
- The legend and color gradients improved multidimensional insights.

Q: What did the geospatial map show?

- A clean outline of **Andhra Pradesh’s new districts**, useful for regional analysis or merging with demographic data.