COMP47970 Visual Exploration Tool Design Document

Student Name: Subhomoy Banerjee

Student Number: 23205360

#### Title:

US Post Offices Analysis

#### Screenshot:

A map of the united states

Description automatically generated

#### Dataset overview:

The Dataset used in this assignment is about the [**US Post Offices**](https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-04-13/post_offices.csv) (Original dataset) taken from the week of [**2021-04-03**](https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-04-13)**.** It contains fields like name of the office, county in which it is located, established year, discontinued year, continuous status which determines if the office operated continuously between the established and discontinued year or not, stamp\_index which determines it scarcity score, duration which calculates the discontinued minus the established year, feature class which is the area in which the office is located, latitude, longitude, elevation, and etc. All the fields are not mentioned in this report as there are many columns out of which some are removed since some of them were redundant and irrelevant.

The updated version of the [**Dataset**](https://raw.githubusercontent.com/subhomoybanerjee/infoviz/main/postal.csv) is pre-processed with [**Python**](https://github.com/subhomoybanerjee/infoviz/blob/main/preprocess.html) pandas (ipynb exported as HTML) used to drop the fields which are not required for the analysis, remove the null values, and remove the established, discontinued, duration columns which are either negative in values or are wrongly recorded in the original dataset, for example some of the years are inputted as a 3 digit or 5 digit quantitative value. Now this step could have been done in vega-lite itself, but the original dataset is too huge and was resulting the online User Interface to crash every time.

#### Design considerations.

**Overall goal:** The Purpose of this analysis was to find the relationships between the geographical locations of US Post Offices and its proximity to surrounding areas, established year, duration of functioning, rarity of postmark (stamp index) and its locations elevation above sea level.

**Geographical Plot:**

The visualization uses the Albers USA projection, which is a composite projection designed for maps of the contiguous United States. The base layer of the plot is a geoshape that represents the countries and the boundaries. This is drawn using data from a TopoJSON file.

The main layer contains circles that represent postal offices. The position of each circle is determined by the longitude and latitude fields (lon and lat). The size of each circle represents the stamp index. This is a nominal field. The color of each circle represents the area (as defined by the gnis\_feature\_class field). The plot includes interactive selections. Users can select specific areas, whether the duration is continuous from the dropdown, and specific stamp indices. The plot will update to show only the data that matches the selected criteria. The plot also includes tooltips that provide additional information when you hover over a circle. The tooltips show the county, stamp index, area, and whether the duration is continuous. The main observation here was that a major portion of the map in between was cleared out when I set the discontinuity to false. Which means that the region which cleared out was an important part of the US postal services and could not shut down in between. But the problem with this plot is that if the dataset is huge, then the whole map will be cluttered with dots leading to overlapping, (which is happening here). However, as I kept applying the filters it became clearer. Which is why a huge dataset is not suitable for plots which are concatenated. This can be corrected by creating a separate plot for the map itself. aggregating the points values works too but information will be lost.

**Bar Chart:**

This is a stacked bar chart. The x-axis represents the duration of postal office activity. This is a nominal field and has been binned into 10 categories for easier visualization. The y-axis represents the count of postal offices. It’s a quantitative field that has been aggregated by count, meaning it shows the number of postal offices that fall into each duration category. The color of the bars represents different areas (as defined by the gnis\_feature\_class field). The chart includes interactive selections. Users can select specific areas, whether the duration is continuous, from the drop-down menu, and specific stamp indices, from the legend. The chart will update to show only the data that matches the selected criterias. The chart also includes tooltips that provide additional information when you hover over a bar. The tooltips show the duration, stamp index, area, count, and whether the duration is continuous. The main observation which I saw is that most of the post offices from all regions were active mainly for a max duration of 0 to 40 years and were near the populated places. But as the Stamp index decreased the duration of operation increased. less stamp index was up and running from duration 60 to 80 years too.

**Arc Plot:**

The arc in the plot represents different areas (as defined by the gnis\_feature\_class field). The angle of each arc (theta encoding) is determined by the count of post offices in each area. This is a quantitative field that has been aggregated by count, meaning it shows the number of post offices that fall into each area category. The arcs are stacked and normalized, meaning the complete circle represents 100% and each arc represents the proportion of post offices, with a focus on whether they were temporarily closed before re-opening or not. The tooltips provide additional context when needed. This plot can be particularly useful for understanding patterns and trends in post office operations across different areas. This plot does not provide anything new and can give similar results when compared to the stacked bar chart. But I included this to see an over result without any complications of visual understanding since pie plots are one of the easiest to understand. Here, I have not applied labels to the plots since they were being overlapped. when the adjacent theta values were small. The only insight I Could get from this that as the stamp index number increased the number of post offices around the populated places decreased and spread to other regions more uniformly.

**Line Plot:**

The x-axis represents the decade when the offices were established. This is an ordinal field, meaning the decades are treated as discrete categories, but with a specific order. The decades are calculated in vega-lite itself. The y-axis represents the mean elevation of the offices in meters. It’s a quantitative field that has been aggregated by mean, meaning it shows the average elevation of the offices established in each decade. The color of each line represents different areas. The opacity of each line is determined by a condition: if the area is selected in the legend, the opacity is set to 1 (fully opaque); otherwise, the opacity is set to 0.1 (mostly transparent). The main and recurrent observation from this plot was that as the decades passed the general elevation of the post offices location increased (they moved to a higher place). the only exception to this trend was the school area post offices which decreased in elevation with decades. But this could also happen if the number of records for less elevation might be less for the later decades which is a why having an unbalanced dataset might be a problem when dealing with averaging. This can be avoided with a better and detailed dataset with enough number of inputs for all the categories.

**Interactions Justification:**

Legend and Dropdown menu Selection is what I went with in all the plots. The stacked bar chart will change it duration categories and show the counts of the Offices, the geoplot will change its points size, colours, opacity and locations, The pie chart will change its normalized Angle values, and finally the line plot will change its aggregated mean elevation peaks and colours of offices location per decade based on the selected area, continuity, and stamp scarcity index. I chose this palette since the area categories had a lot of values which in my opinion was best distinguishable with the “rainbow” colour scheme. However, the main disadvantage of all the plots alongside with the interactivity is that the dataset is huge even after the preprocessing stage. It is taking a lot of time for vega-lite to render the plots, especially after a specific area or stamp index is chosen, which is why I have chosen not include more interactivity, Also the dataset has important entries which cannot be dropped for the sake of faster rendering. Other than that, the geoplots can be edited to form contour plots to show the elevations more precisely and in a detailed manner.