Analysis of wage data CO

https://data.colorado.gov/Labor-Employment/Employment-Wages-in-Colorado/busm-qa5b

Lyric

Taste profile

Concepts to explore:

Part 3

1. **What is the problem you want to solve?**

   How have wages changed as Denver's population has increased. How have different regions been effected. What industries have experienced wage growth in a rapidly growing city

2. **Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?**
3. **What data are you going to use for this? How will you acquire this data?**

   https://data.colorado.gov/Labor-Employment/Employment-Wages-in-Colorado/busm-qa5b

4. **In brief, outline your approach to solving this problem (knowing that this might change later).**

   Start to form timelines by different region and become more specific by sector

5. **What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.**

Part 4

Now that you have a basic ideas of the various data wrangling steps and techniques available, let's apply it to your Capstone Project! By now, you probably have a data set in mind for your project (If you don't have a data set yet, come back to this assignment once you have one). Apply some of the data wrangling techniques you have learned to this data set.

Submission: Create a short document (1-2 pages) in your github describing the data wrangling steps that you undertook to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them? This document will eventually become part of your milestone report.

Part 4 data wrangling

The first step I did in the data cleaning portion was to find a list of columns that contained blanks. I used a function to return true if there is a blank in any row and received an output telling me that of the 40 columns 15 columns contained blank fields

Those columns are

`mean`	Mean wage for the occupation.
`entrywg`	Entry level wage for the occupation, mean of the first third (ALC definition).
`experience`	Experienced level wage for the occupation, mean of upper two thirds (ALC definition).
`pct10`	Wage at tenth percentile.
`pct25`	Wage at 25th percentile.
`median`	Median wage of the occupation; also the wage at fiftieth percentile.
`pct75`	Wage at 75th percentile.
`pct90`	Wage at 90th percentile.
`udpct`	User defined percentile.
`udpctwage`	Wage at user defined percentile
`udrnglopct`	Low percentile for user defined range.
`udrnghipct`	High percentile for user defined range.
`udrngmean`	mean wage for user defined range.
`wpctrelerr`	Relative percent error on wage.
`panelcode`	Reference panel code (yyyymm)

The next step I wanted to check was to see what percentage of each dataset is missing.

```
mean has 1742 null values, 2.0 percent is missing
entrywg has 1912 null values, 2.0 percent is missing
experience has 1974 null values, 2.0 percent is missing
pct10 has 1762 null values, 2.0 percent is missing
pct25 has 1852 null values, 2.0 percent is missing
median has 2083 null values, 2.0 percent is missing
pct75 has 11082 null values, 13.0 percent is missing
pct90 has 11958 null values, 14.0 percent is missing
udpct has 56179 null values, 67.0 percent is missing
udpctwage has 74887 null values, 90.0 percent is missing
udrnglopct has 56179 null values, 67.0 percent is missing
udrnghipct has 56179 null values, 67.0 percent is missing
udrngmean has 74887 null values, 90.0 percent is missing
wpctrelerr has 9091 null values, 11.0 percent is missing
panelcode has 35651 null values, 43.0 percent is missing
```

Next I wanted to find duplicate rows so I ran a basic function looking for duplicates and found the result to be zero. However, I noticed that there is a bunch of summary data mixed in with specific metropolitan data. I found that the column areatyname specifies a few categories

```
['State' 'Metropolitan Statistical Area' 'Balance of State (pre-2015)'
 'Balance of State']
```

We really only care about finding the information of metropolitan statistical area as that contains the specifics of each city we are exploring. This way we can reduce redundancies in the data that we get more accurate information. I also believe that specifying it by city will reduce a lot of the gaps in the data that we specified was missing above.

We find first that it reduced the dataset to 50,612 rows (removed 40% of data). Now let's rerun the for loop to see how the null values change

```
mean has 1013 null values, 1.0 percent is missing.
entrywg has 1121 null values, 1.0 percent is missing.
experience has 1155 null values, 1.0 percent is missing.
pct10 has 1029 null values, 1.0 percent is missing.
pct25 has 1089 null values, 1.0 percent is missing.
median has 1220 null values, 1.0 percent is missing.
pct75 has 6522 null values, 8.0 percent is missing.
pct90 has 7121 null values, 9.0 percent is missing.
udpct has 33539 null values, 40.0 percent is missing.
udpctwage has 45191 null values, 54.0 percent is missing.
udrnglopct has 33539 null values, 40.0 percent is missing.
udrnghipct has 33539 null values, 40.0 percent is missing.
udrngmean has 45191 null values, 54.0 percent is missing.
wpctrelerr has 5453 null values, 7.0 percent is missing.
panelcode has 21720 null values, 26.0 percent is missing.
```

In order to remove redundancies we should look at the way the wage data is classified. We can find that the wages are defined as either hourly or annual wage. I have a feeling that the data is split between the two.

Let’s go ahead and check.

It looks like the data is split. Hourly wage=25146
Annual salary= 25466

For the sake of eliminating redundacnies I am going to keep only the annual salary columns.

This has drastically reduced the amount of null values especially in key columns like mean and median

```
mean has 50 null values, 0.0 percent is missing.
entrywg has 104 null values, 0.0 percent is missing.
experience has 121 null values, 0.0 percent is missing.
pct10 has 58 null values, 0.0 percent is missing.
pct25 has 88 null values, 0.0 percent is missing.
median has 154 null values, 0.0 percent is missing.
pct75 has 2860 null values, 3.0 percent is missing.
pct90 has 3167 null values, 4.0 percent is missing.
udpct has 16883 null values, 20.0 percent is missing.
udpctwage has 22709 null values, 27.0 percent is missing.
udrnglopct has 16883 null values, 20.0 percent is missing.
udrnghipct has 16883 null values, 20.0 percent is missing.
udrngmean has 22709 null values, 27.0 percent is missing.
wpctrelerr has 2781 null values, 3.0 percent is missing.
panelcode has 11020 null values, 13.0 percent is missing.
```

It looks like there has been a significant reduction in the null values across all columns with null values. Next let’s explore outliers in the mean column. It looked like everything was within a reasonable measure. I explored which occupations were listed as outliers on the upper end and found Actors and various doctors. We can assume this means that any outliers are within reason and not due to entry errors.

How does one go about creating a data story? You have some pointers from the material you've just gone through, but they're probably a bit on the abstract side when you're just getting started. Also, storytelling is an art, so you have to get your imagination going. Here are some pointers to get those creative juices flowing. In the following sections we will work step-by-step to create your first Data Story

1. Pick a dataset - ideally the dataset for your Capstone. If for some reason you want to do this on a different data set, you can find one on Mode Analytics or Google's public data sets directory, or pick another one you like from elsewhere.
2. Get going by asking the following questions and looking for the answers with some code and plots:
    1. Can you count something interesting?
    2. Can you find some trends (high, low, increase, decrease, anomalies)?
    3. Can you make a bar plot or a histogram?
    4. Can you compare two related quantities?
    5. Can you make a scatterplot?
    6. Can you make a time-series plot?
3. Having made these plots, what are some insights you get from them? Do you see any correlations? Is there a hypothesis you would like to investigate further? What other questions do they lead you to ask?
4. By now you've asked a bunch of questions, and found some neat insights. Is there an interesting narrative, a way of presenting the insights using text and plots from the above, that tells a compelling story? As you work out this story, what are some other trends/relationships you think will make it more complete?

Submission Requirements: Submit links to a Github repository containing a Jupyter notebook. The notebook should contain:

1. The questions you asked
2. The trends you investigated
3. The resulting visualizations and conclusions.

In case the data set is too large to commit to Github, please include a link to the data set inside the Jupyter notebook.