CS/STAT 24200: Project 1

Due date: Thursday Mar 26, 11:59pm

You must code up and submit this project as a single python Jupyter notebook: do not add any of the data files, and do not zip it. Your notebook should contain both the code that you used for analysis as well as your answers to the questions below. Use plots to help justify answers to questions: do not make claims without any evidence. Submit this to blackboard.

We will use three datasets,

- `fatalities.csv`, a dataset of US traffic fatalities for the lower 48 US states (i.e., excluding Alaska and Hawaii) annually for 1982 through 1988. This dataset also has state-level yearly information, such as population size, beer tax, employment level, among many other. See `https://rdrr.io/cran/AER/man/Fatalities.html` for a description of the covariates, we will mostly focus on `fatal, afatal, popl, beertax, year`.
- `states`, a dataset of state-level geographic information to mark out state boundaries.
- `state_abbrv`, a dataset mapping state names to two-letter abbreviations.

In this assignment, you will transform, explore and analyze the first dataset.

1 Simple visualizations (10 pts)

Read in the dataset `fatalities.csv` and plot scatterplots `afatal` vs `popl`, `afatal` vs `miles` and `afatal` vs `fatal`. Make sure to label the plots appropriately.

Comment on your results. Do you see any batches of outliers. Do they correspond to a particular state or a particular year? By adding color/shapes to your plots, investigate this. Which were the top 5 and bottom 5 states in terms of alcohol-related fatalities. What is the overall trend across time for alcohol-related fatalities?

2 Normalizing Data (10 pts)

To correct for population effects, add two new columns `fatal1000, afatal1000`. These correspond respectively to the number of traffic fatalities per thousand people, and alcohol-related traffic fatalities per thousand people. Again, visualize the relationship between these variables, accounting for state and year info (in the same or different plots). What were the top 5 and bottom 5 states in terms of `afatal1000`? And what is the trend across time?

3 Visualizing maps (10 pts)

Since there are about 50 states, understanding patterns across states using different colors or shapes is not easy. Instead, we will plot this on a map of ths US. As a first step, load the dataset `states.csv`, and plot it as we did in Lab 5 (i.e. use `geom_path` after grouping by `group`). You should see the state-boundaries.

Next add a layer `geom_polygon`. It should take the same aesthetics as `geom_path`, additionally, set `fill='region'`. Now, each state in the US map should be colored differently. Note, that if we had another column (e.g. `afatal1000`), then setting `fill='afatal1000'` would color each state according to the alcohol fatality-rate for that state. We will do this next.

4 Merging data (10 pts)

Write a function that takes a dataframe `ip_df` as input, along with a column name `col_name`. The function merges values from this column to `states` with each element of `col_name` passed to rows in `state` of the appropriate state. Since `states` uses full state names, and `fatalities` uses 2-letter abbreviations, use the `state_abbrv` dataset. Note that `states` has `'district of columbia'` as one of its regions, and does not include Alaska and Hawaii, so you should drop these from both datasets.

5 Correlation and Scatterplots (20 pts)

Use the above function to plot a map of the US, filling each state with its alcohol-related fatality rate averaged across years. Create another plot with two facets, one showing alcohol-related fatality rate in 1982 and the other in 1988 across all states. Finally, create a plot showing the change in alcohol-related fatality rate between 1988 and 1982 across all states. Comment on your results. Which state had the biggest drop? Which had the smallest drop/biggest increase? To each of these plots add a `geom_text` layer giving the 2-letter state abbreviation.
Note: For these plots, use a colormap that shifts from one color to another (e.g. blue is low and red is high). Look at the documentation of `scale_fill_continuous`*.*

6 Effect of `beertax` on alcohol fatality rate (20pts)

Create a plot with two facets, both having `year` on the x-axis. The left has `afatal1000` on the y-axis, the right has `beertax`. By default `facet_wrap` and `facet_grid` use the same y-axis for both facets, to avoid this, use `facet_wrap` with the option `scales='free_y'`. (see the documentation)

Comment on your results. Fit a linear regression model with `beertax` as input and `afatal1000` as output. Report and interpret your parameters. You should get a negative slope. Is this surprising?

7 Effect of beertax on alcohol fatality rate (contd) (20pts)

To better understand the effect of beertax, we will account for the fact that different states might have different alcohol-related behavior. For instance, Utah might have low alcohol tax as well as low alcohol-related fatalities.

To do this, allow each state to have its own intercept, while sharing a single slope parameter. This keeps the effect of changing beer-tax across states; however setting the beer-tax to 0 results in different fatalities in different states. You can do this either by creating new columns one-hot encoding each state, or by just passing the state column in your linear regression formula: since the states are strings, package smf will automatically one-hot encode them.

How many parameters does this model have? How many datapoints do you have to set these parameters?

Having fit the model, create a map-plot filling each state with its associated intercept. Also report on the coefficient on beer_tax.

8 Accounting for time (20pts)

In addition to fluctuations across states, one can have fluctuations with year. E.g. maybe the year 1982 witnessed a spike in gas prices that caused a drop in vehicle accidents across all states. To account for this, we will introduce into the earlier model 7 more variables, an indicator variable for each year.

Fit a linear regression model with this additional co-efficients, and create a bar plot showing them. Do you see any significant fluctuations across years. Once again, plot the state-level coefficient (like the previous question), and report the slope of beer_tax. Are these results different from the previous question?