Data Visualisation Assignment 1

## Name: Maurício de Oliveira Reis

## ID: D21125621

## Programme: TU256/1

## Class: Data Visualisation SPEC9995: 2022-23 - Part Time

## Dataset Adult Data

This dataset contains information extracted from the 1994 US Census database and is commonly utilised to create prediction tasks to determine whether a person makes over 50K a year. For this work, I’ll be using just the adult.data dataset. More info about this dataset can be found at <https://archive.ics.uci.edu/ml/datasets/adult>.

#### Data Clearning

Data cleaning was done utilising python and pandas, the script utilised to do that is attached to this work with the name dataCleaning.ipynb. The only clearing and formatting done by the script consists in putting names on the columns, identifying and removing lines with NA values and changing the file name to add a .csv in the end, both files adult.data (the original) and adult.data.csv (the ocnverted) are included into this submission

As there were very few NA values, the rows from the dataset that had one or more NAs got removed. Below you can see the detail of each row removed outputted by the cleaning script:

**In a total of 32561**

workclass has 1836 NAs which is 5.64% of its rows.

occupation has 1843 NAs which is 5.66% of its rows.

native-country has 583 NAs which is 1.79% of its rows.

**There are 30162 rows left after removing NAs**

#### Variables Description

* salary: Categorical. Salary is the target variable in this dataset and its values represent salary above or below 5k a year.
* age: Numerical. Age represents a particular person’s age during the census date, on this dataset it varies between 17 and 90 years old.
* workcalss: Categorical. Represents the work class that the census respondent belong to. The workclasses on this dataset are: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
* fnlwgt: Numeric. Can be used as ID as is unique for every row, this is a number that represents a group of people in the population that is similar. This number aggregates all the other attributes therefore, people with numbers close to each other have similar attributes. This number goes from 13769 to 1484705.
* education: Categorical. Refers to what education level the census respondent achieved. The education levels on this dataset are: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
* education-num: Numerical. Similar to education but represented in numbers from 1 to 16.
* marital-status: Categorical. Represents the marital status of the census respondent. The marital status on this dataset are: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
* occupation: Categorical. Represents what this person works on. The occupations listed on this dataset are: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
* relationship: Categorical. Describes your family relationship status. The relationships listed on this dataset are: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
* race: Categorical. Describes the ethical race the census respondent belongs to . The races on this dataset are: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
* sex: Categorical. This dataset only considers the sex: Male or Female.
* capital-gain: Numeric. Represents how much money the census respondent made in capital gain in the last fiscal year before the census. The values vary between 0 and 99999.
* capital-loss: Numeric. Represents how much money the census respondent lost in capital losss in the last fiscal year before the census. The values vary between 0 and 4356.
* hours-per-week: Numeric. Represents how many hours the census respondent works per week. Varies between 1 and 99.
* native-country: Categorical. Country where the person was born. The countries listed on this dataset are: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

## Intended Audience

This content will be produced for a left leaning youtube channel about politics and economics. Its owner wants to feature its content both as shared image posts on youtube as parts of a video series about the history of income inequality in the US. Its viewers are from diverse backgrounds but the guidelines given to me is that I should expect people with some understanding about economics and statistics. People who watch this channel usually have a high interest in those areas even though, for many, their profession is not directly related to those areas. They are adults between 25 and 40 years old. Mostly from US.

## Tableau Workbook

The Tableau Workbook for this work can be found on <https://public.tableau.com/views/DataVisualisationAssignment1_16785759322720/ExplanatoryDashboard?:language=en-GB&publish=yes&:display_count=n&:origin=viz_share_link>

## Data Exploration

This dataset has 15 fields and 30162 rows. I’ve decided to start my exploration on analysing salary vs capital gain/loss and Race vs Salary and then mix both analysis.

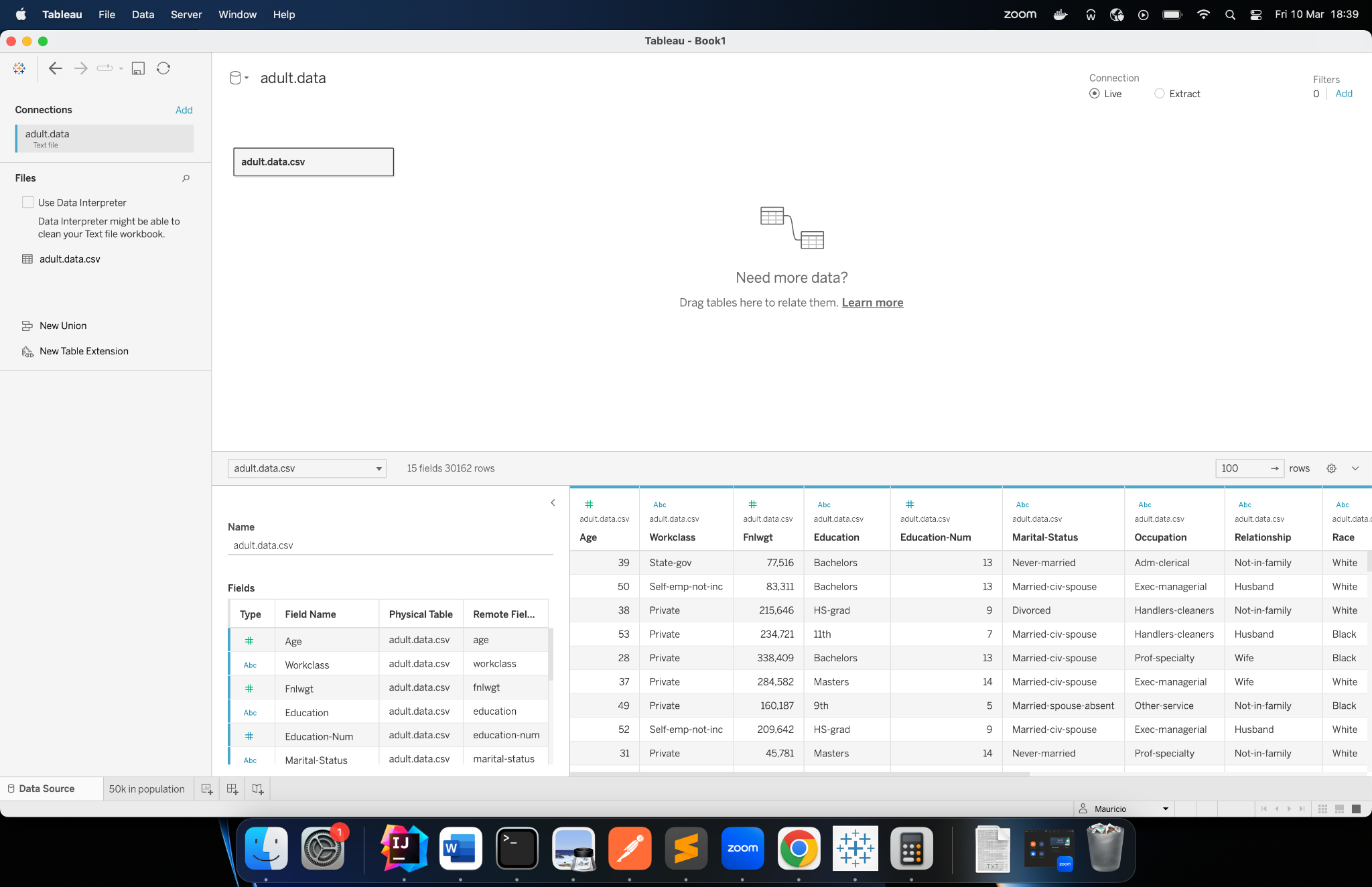


Image 2: Data Source View of adult.data dataset

Before starting working on that, is important to see if there isn’t any person with both capital gain and loss in the dataset which would be a data inconsistency, the chart below shows that nobody has both capital gain and loss

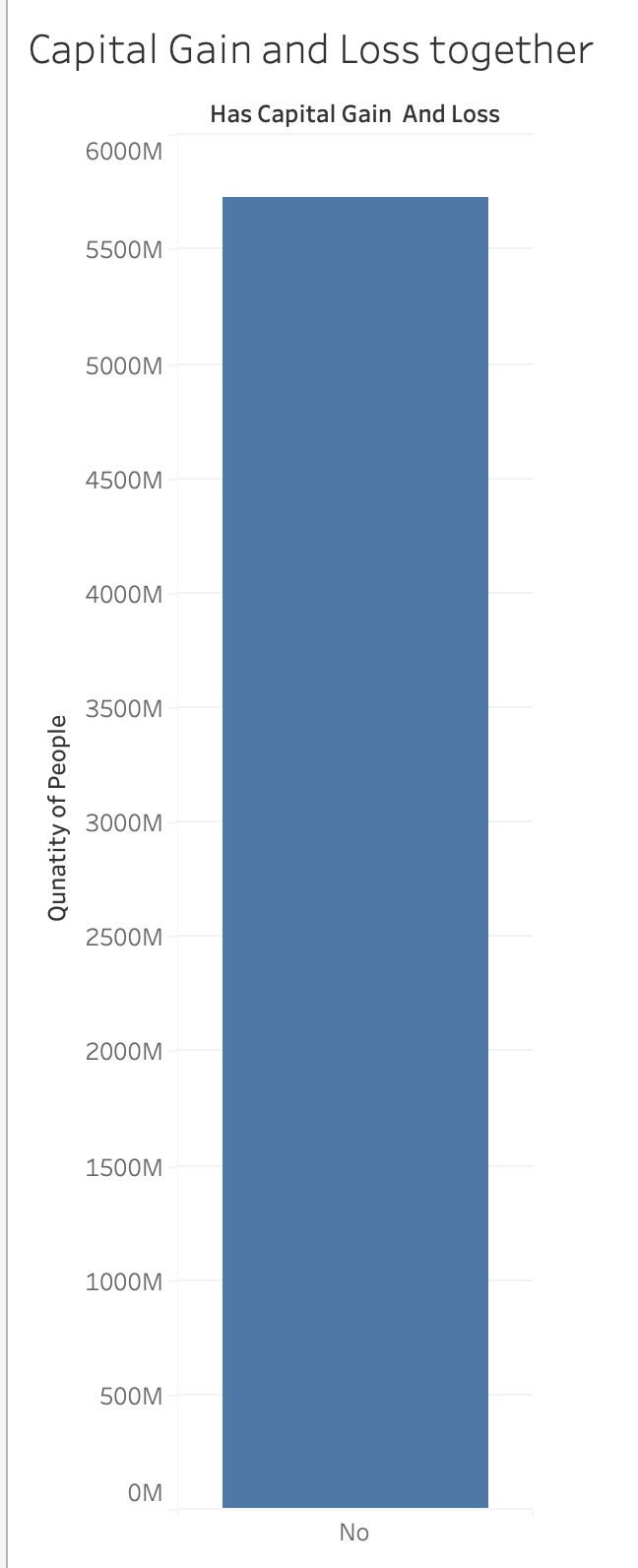


Image 3: Capital Gain and Loss Together

The proportion of people that makes more than 50k a year is much smaller than who make more than that, it’s just about ¼ of the population

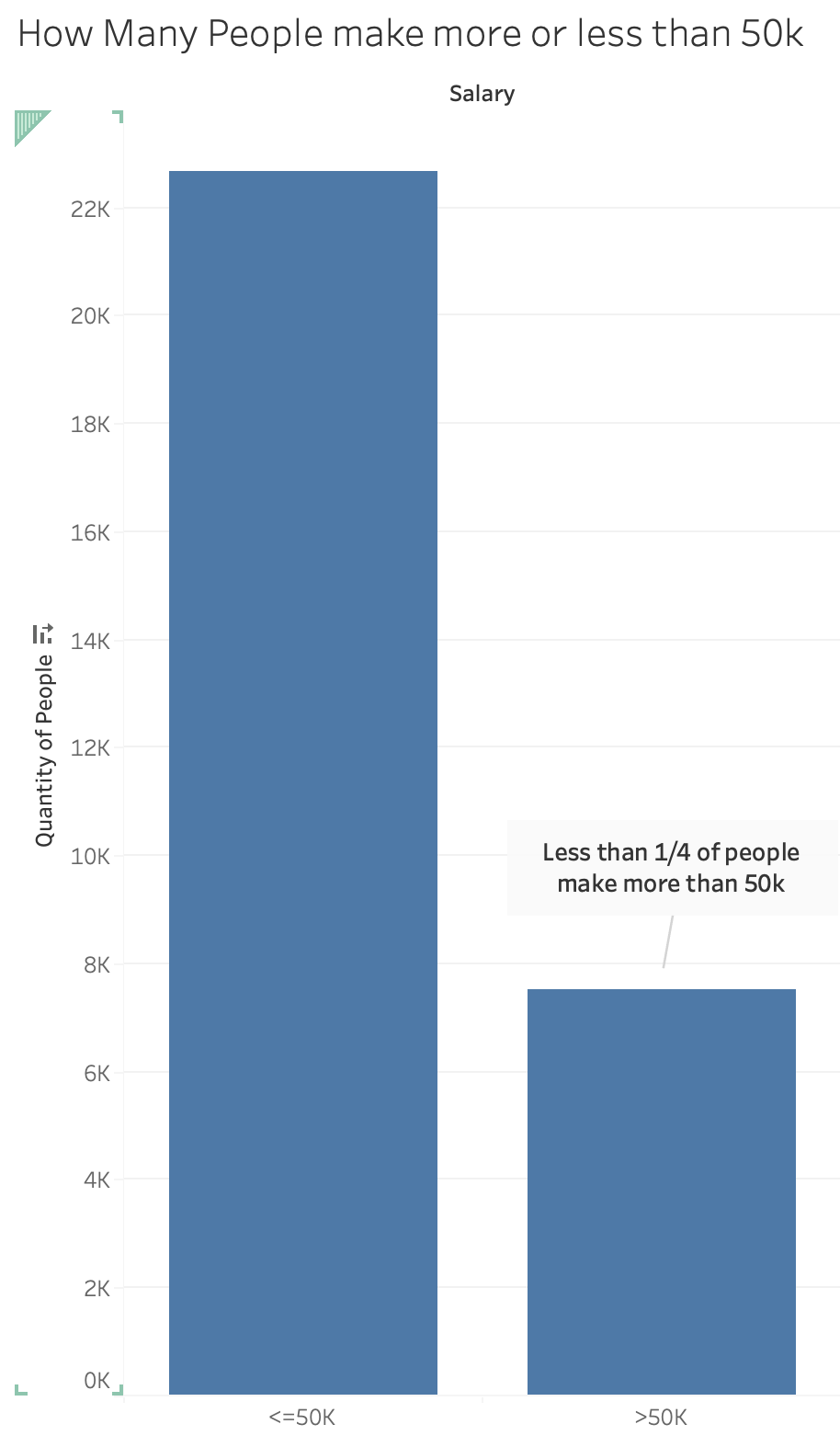


Image 4: How Many People make more or less than 50k

But, if we look at the total capital gain for each one of the two groups above, the ¾ of the less than 50k earners, amount for only 10.24% of all the capital gain.

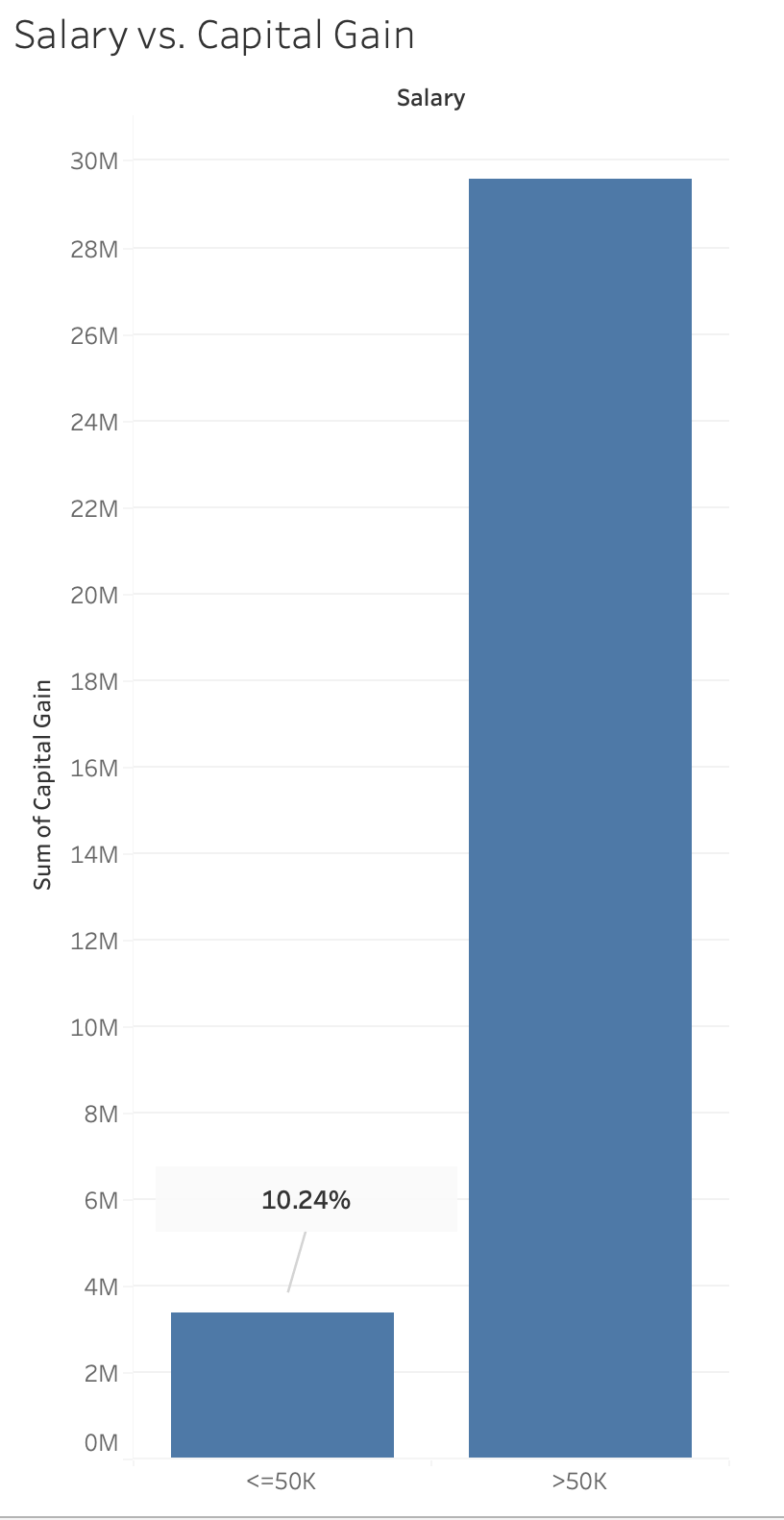


Image 5: Salary vs. Capital Gain

But we also see that the the capital loss is more concentrated in the group who makes more than 50k:

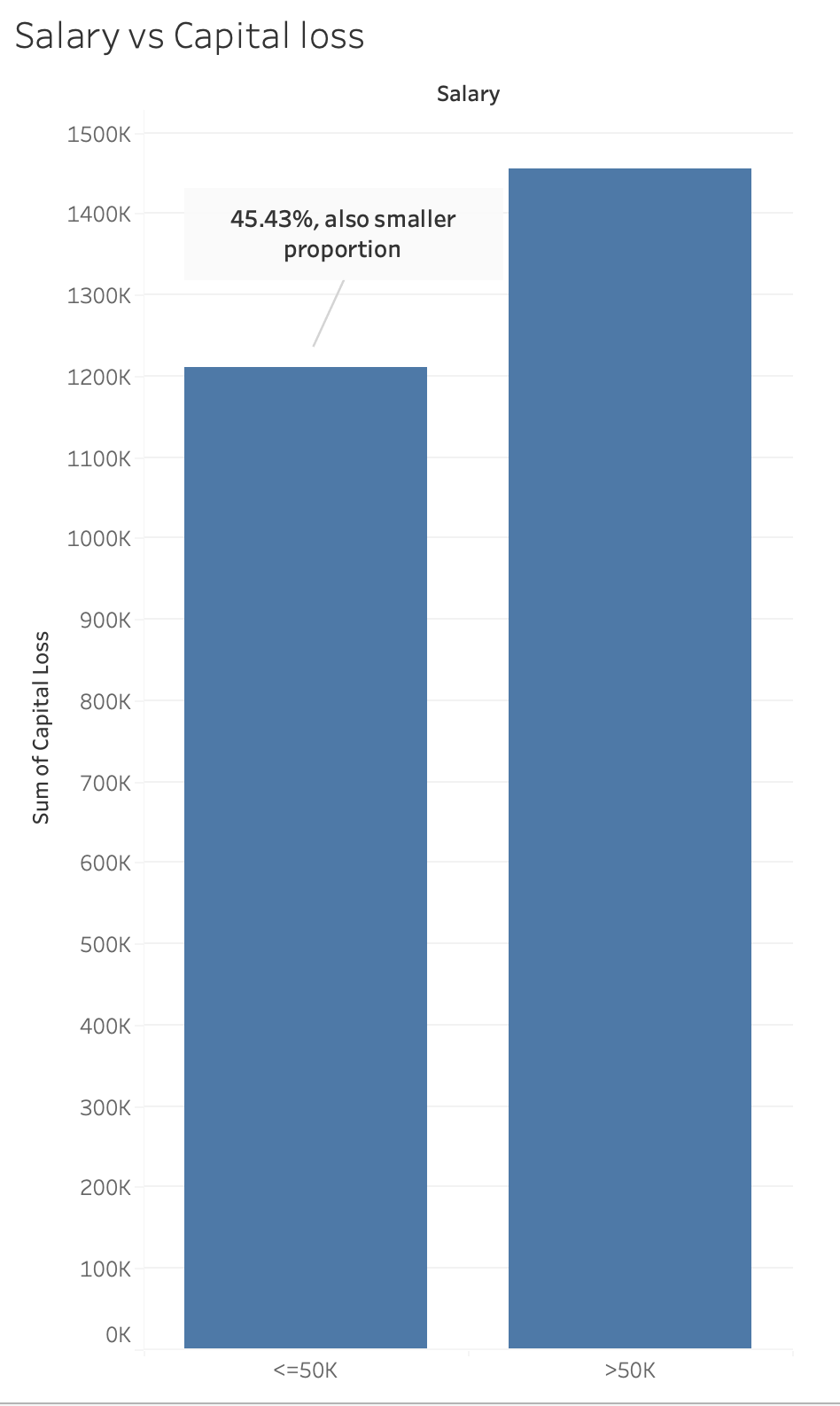
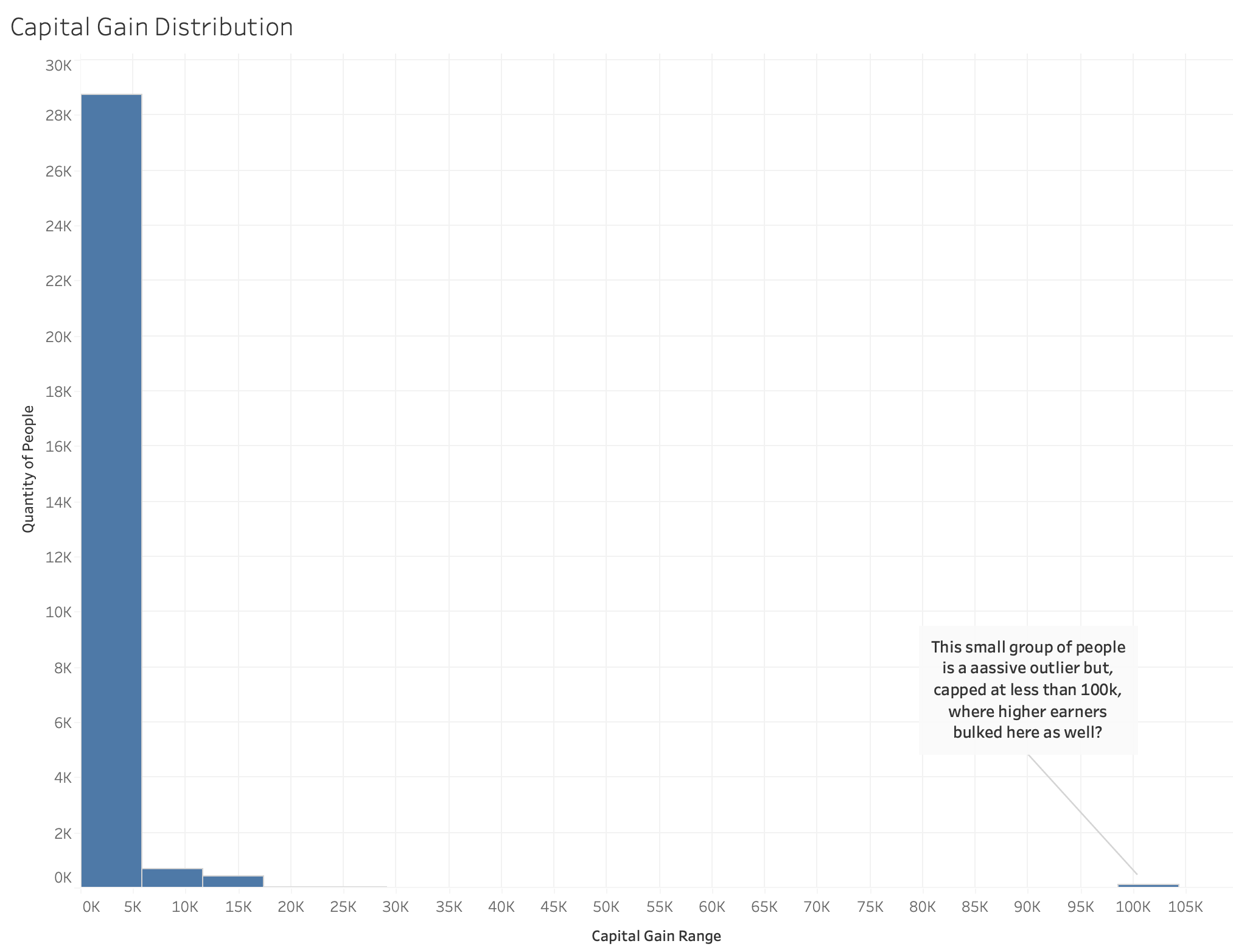


Image 6: Salary vs. Capital Loss

That can be due to a few people making very high gains or losses and skewing the results. But below we see that gains are capped at less the 100k and there are very few individuals who make this amount and are huge outliers, most make less than 17k on gains:

Image 7: Capital Gain Distribution

Also, on that small top capital gain bin, everyone makes more than 50k

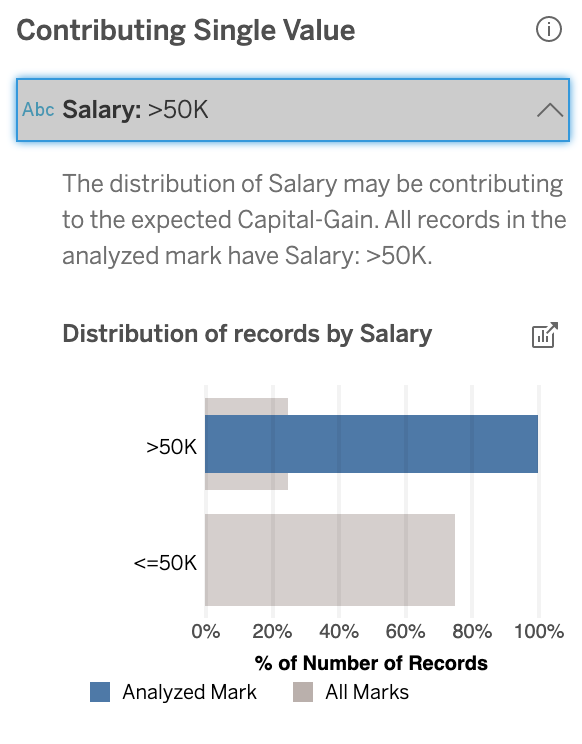


Image 8: Insights about Salary on highest bin of Capital Gain

In capital loss most of the population is concentrated near zero or zero too. It looks that there is a separate normal distribution within this distribution with 4 bins between 1390 and 2502 in capital loss.

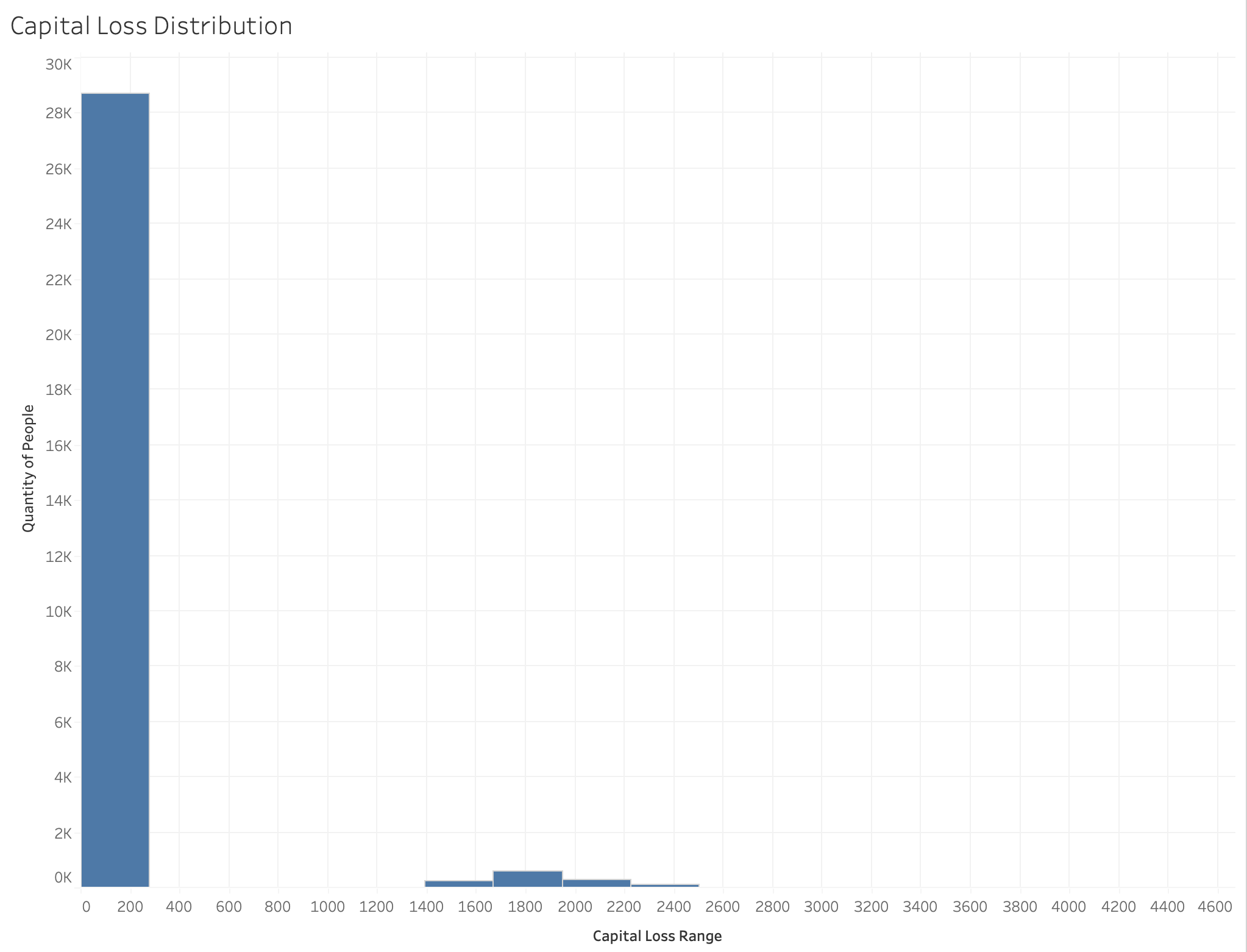


Image 9: Capital Loss Distribution

Now let’s see the proportions of people who have any capital gain/loss at all vs salary:

Most people don’t have capital gains but, for those who have, the <50k salary group, even representing less than ¼ of the population, here represents more than 63% or people who has capital gain

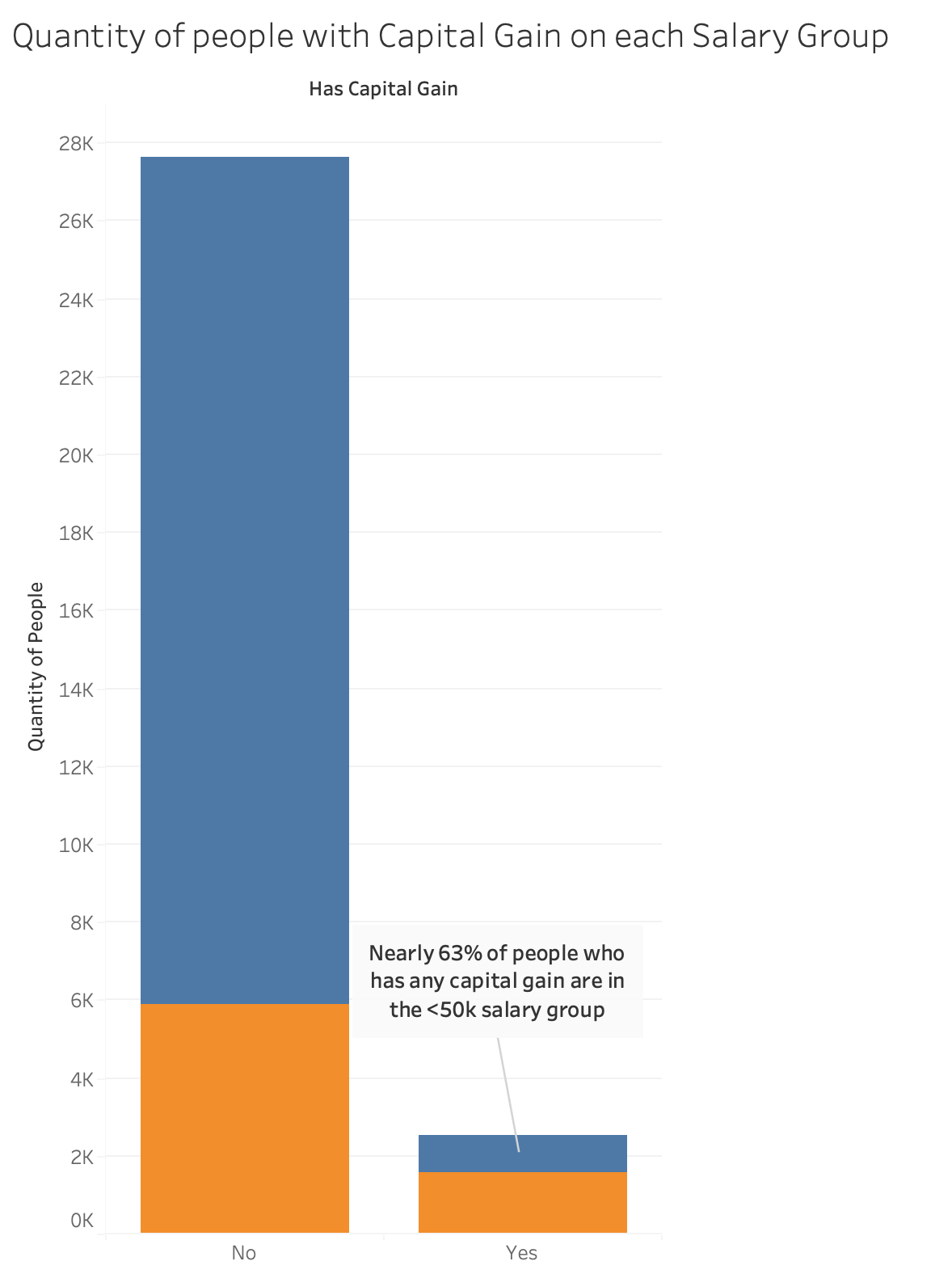


Image 10: Quantity of People with Capital Gain on each Salary Group

On Capital loss the proportion of people who have loss is smaller than in gain but, the majority who have it also make more than 50k a year.

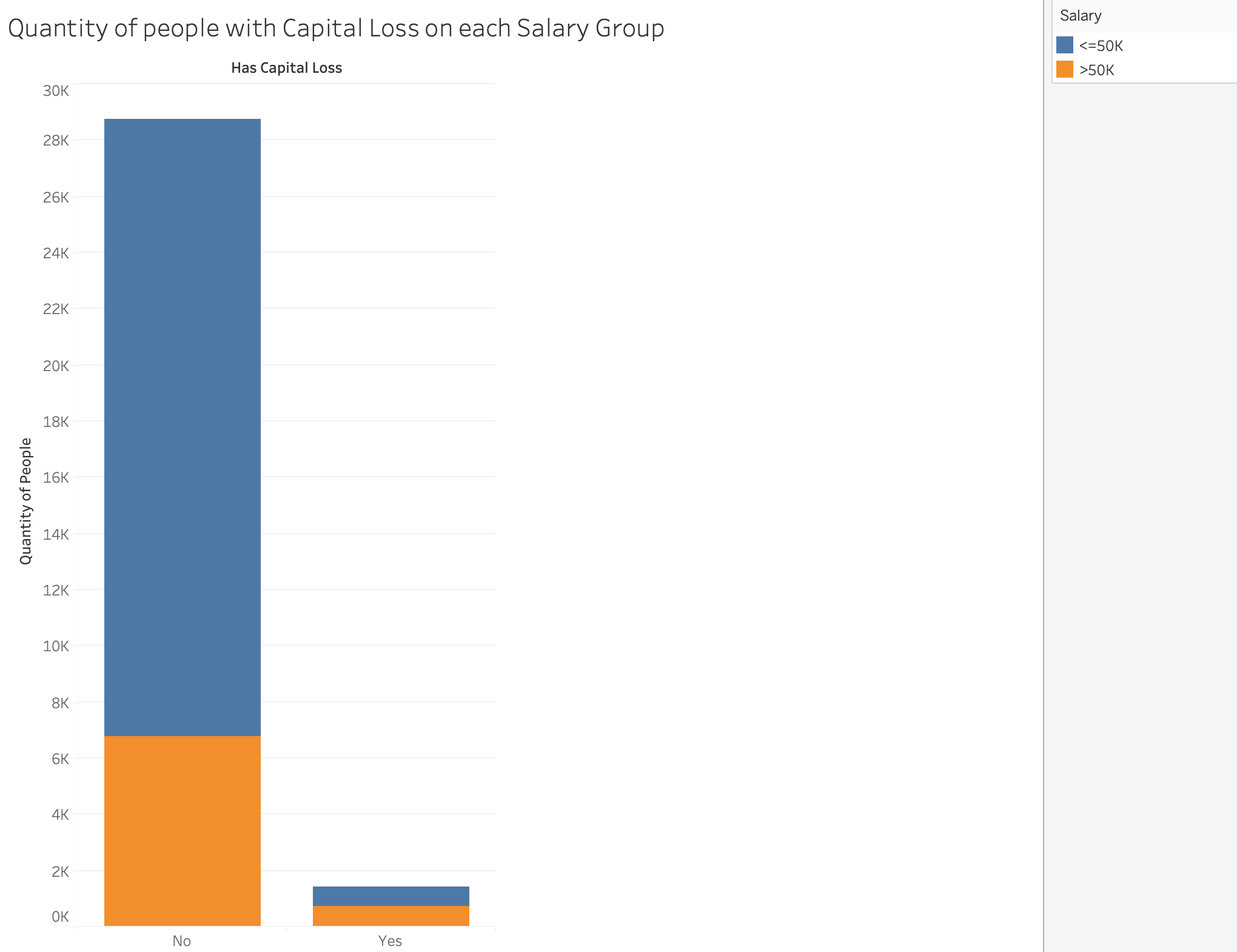


Image 11: Quantity of people with Capital Loss on Each Salary Group

For a person to have either capital gains or losses, this person needs to have assets so, we can make a chart which captures the proportions of people making more or less than 50k based on having either capital gain or loss, we already know that having assets that generate either capital gains or losses is correlated to <50k salary but this chart groups both to give one single view:

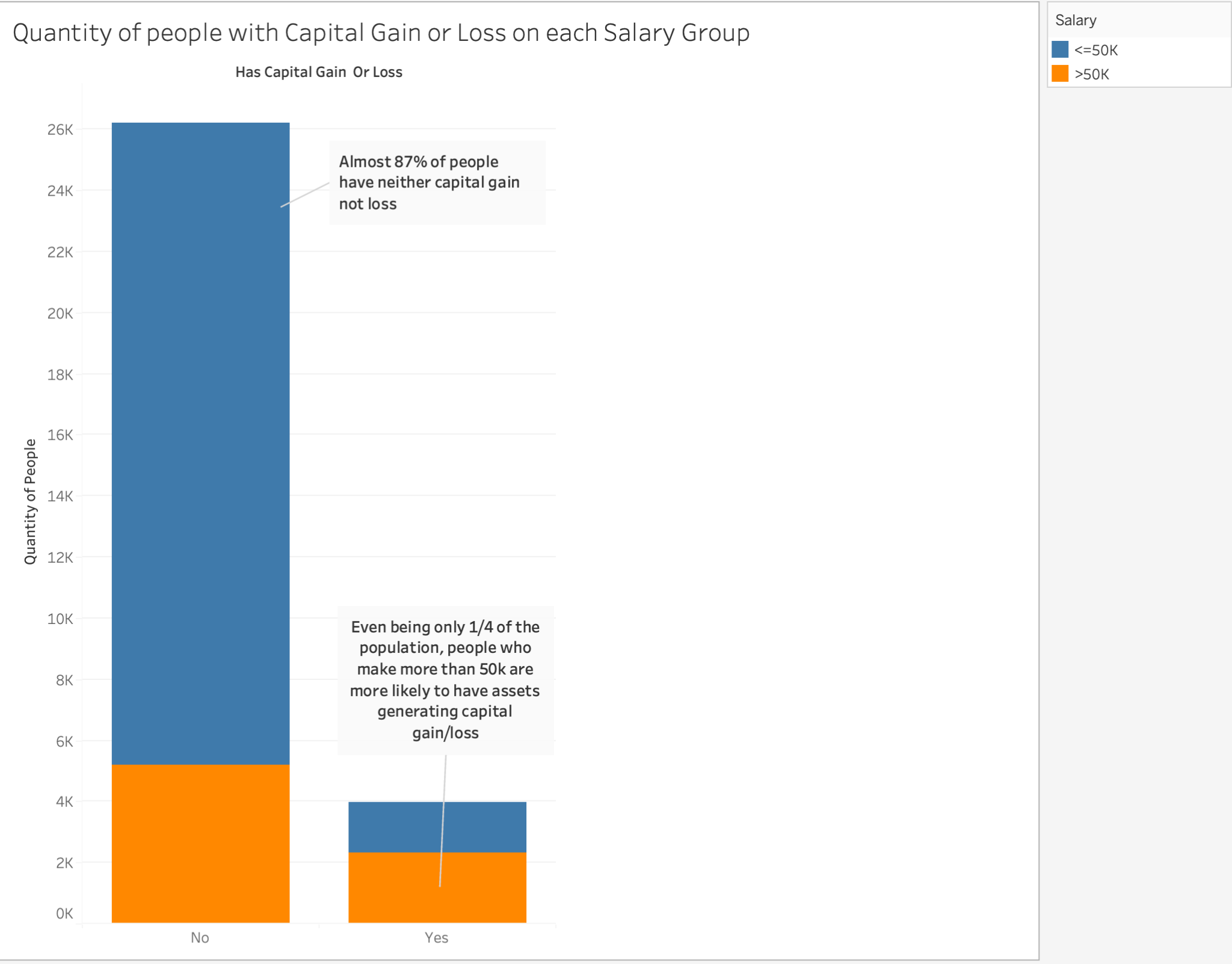
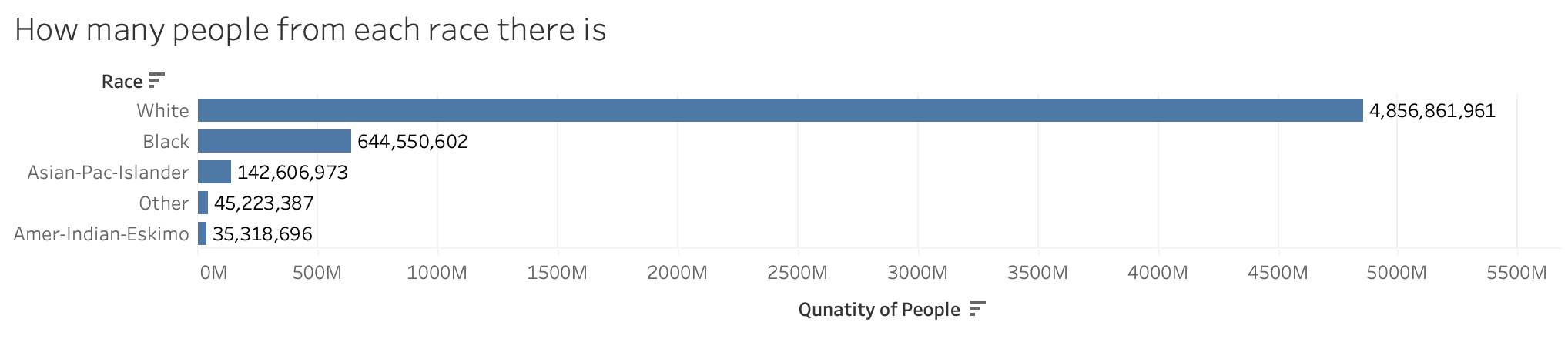


Image 12: Quantity of people with Capital Gain or Loss on each Salary Group

On Race, we can see that most people on this dataset are white:

Image 13: How many people from each race there is

We can also see that Asian-Pac-Islander are closely followed by Whites on who has higher percentage of people making more than 50k

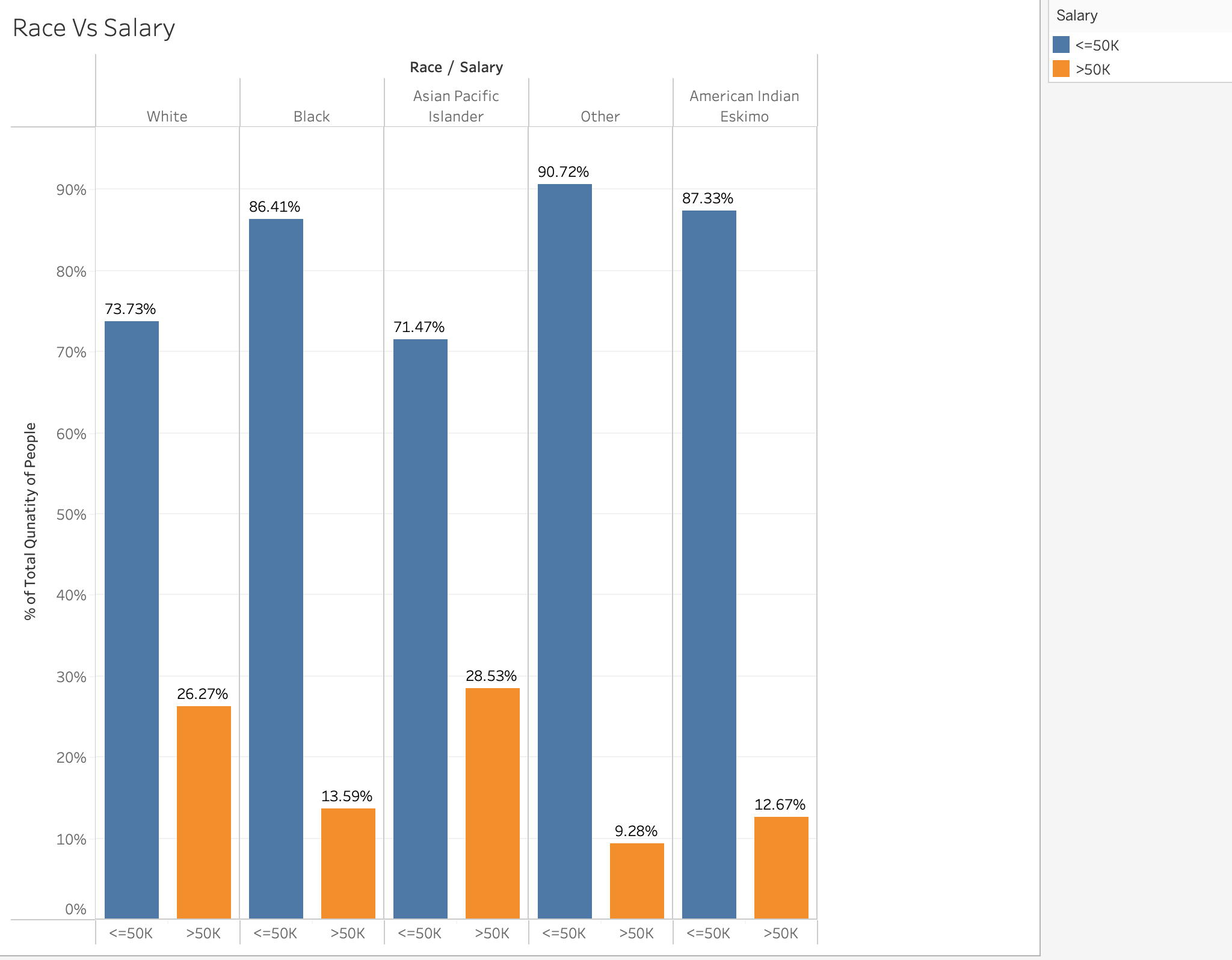


Image 14: Race vs Salary

Putting all together in a Dashboard (Image 15) I could use Race as a filter and notice things like for example the Eskimo Race don’t have any individual with capital gains on that top bracket, that’s the only Race that doesn’t have it. The dashboard allows to use other races filters and also a salary filter to compare the aspects presents on the 6 charts present on it.

## Exploratory Dashboard

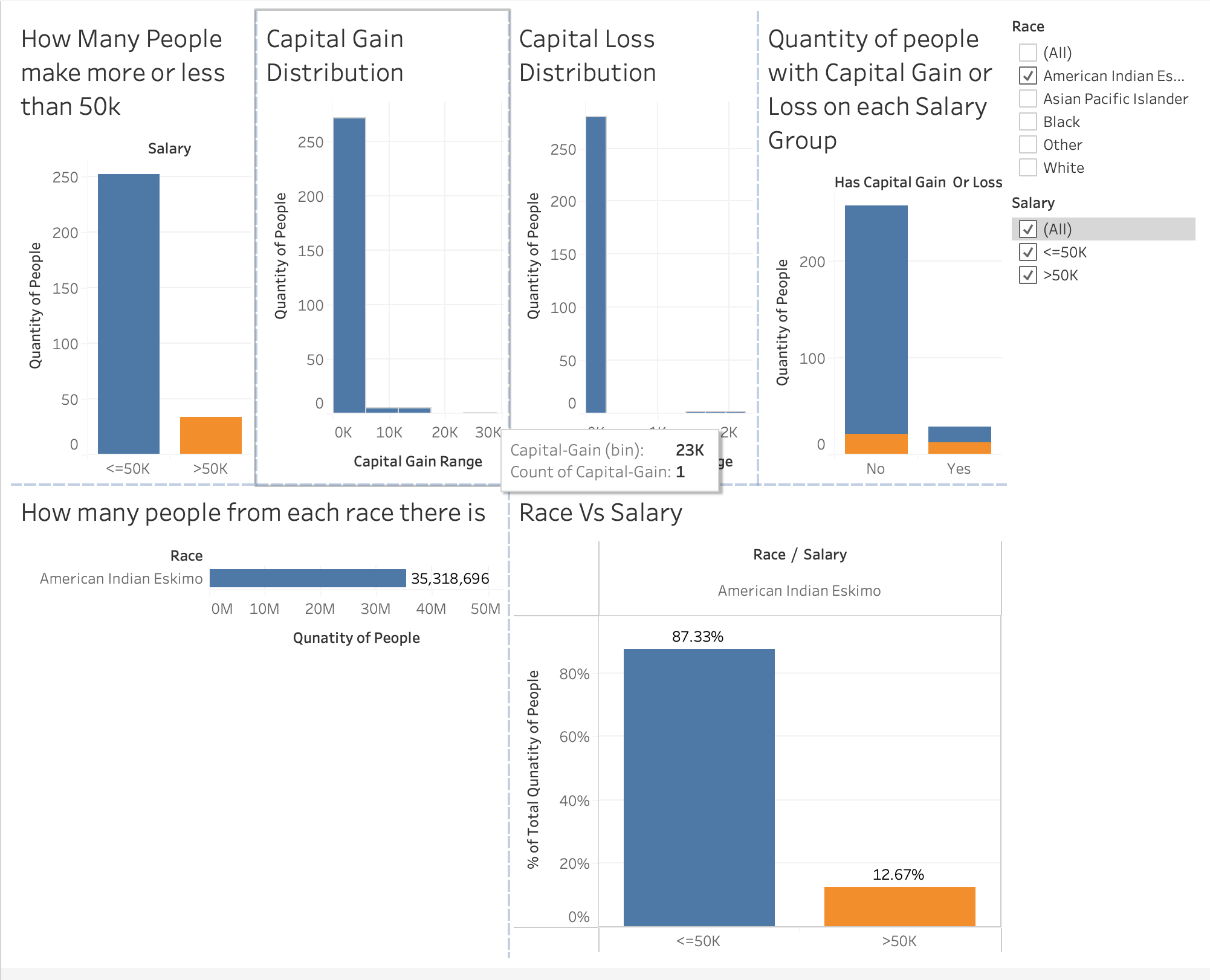


Image 15: Exploratory dashboard filtered by American Indian Eskimo

## Explanatory Dashboard with Insights

On the Explanatory Dashboard, we can see 3 charts that tell a story of the possible relationship between income and wealth inequality where people who already make higher salaries also have assets to sell. This inequality can also be seen looking into races where Whites and Asian Pacific ones are the more favoured.

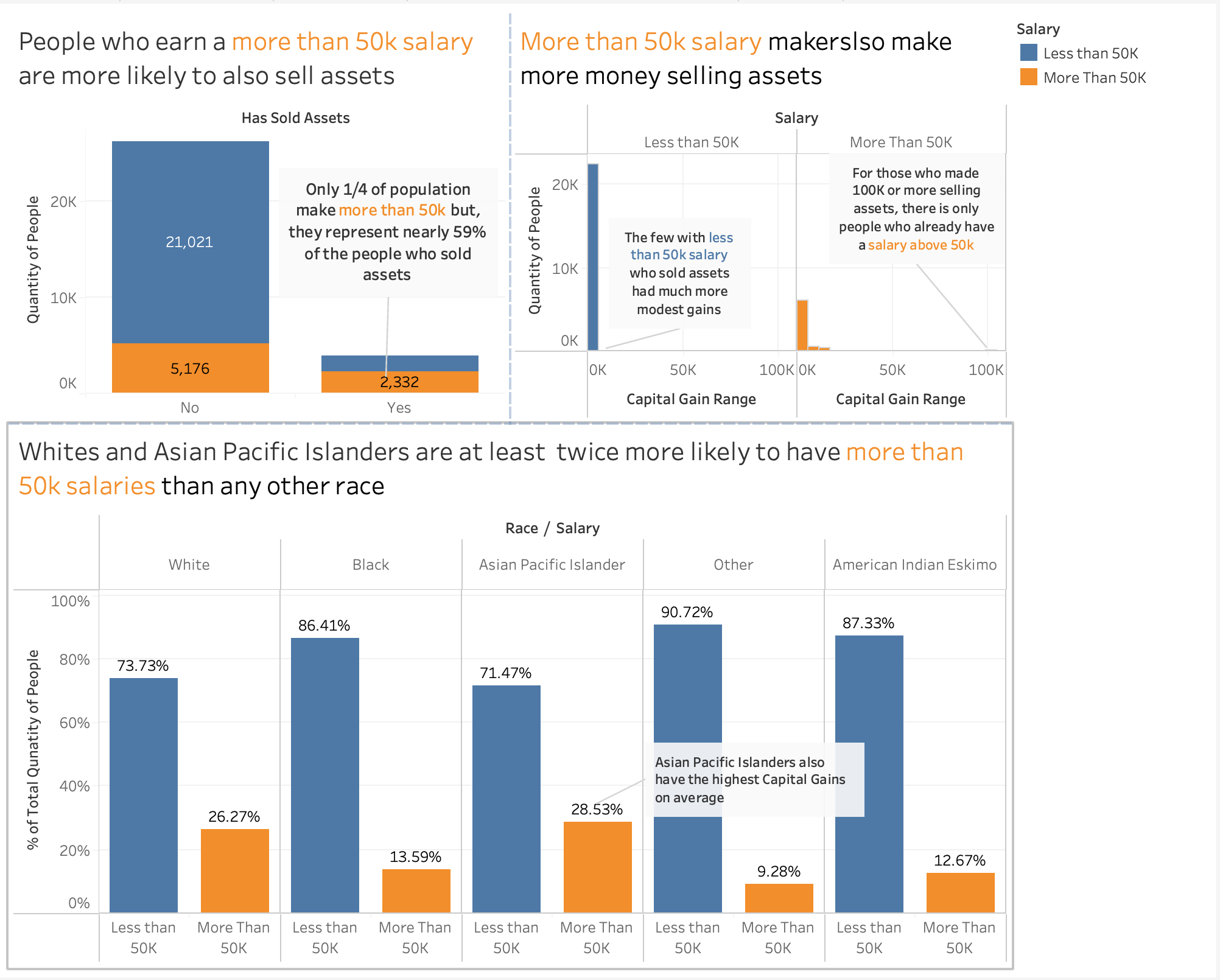


Image 16: Explanatory Dashboard

## Iterations

1. I had this work done for Data Mining that was a small python script to clean data and work on Orange Data Mining. Due to Orange’s limitations, I couldn’t do a great analysis of that work. I probably can revisit that work, reuse the python script and do the Exploration on Tableau to see how much better I’ll do it, especially given that we worked with nearly identical dataset in one of our labs and I already felt potential on that work.

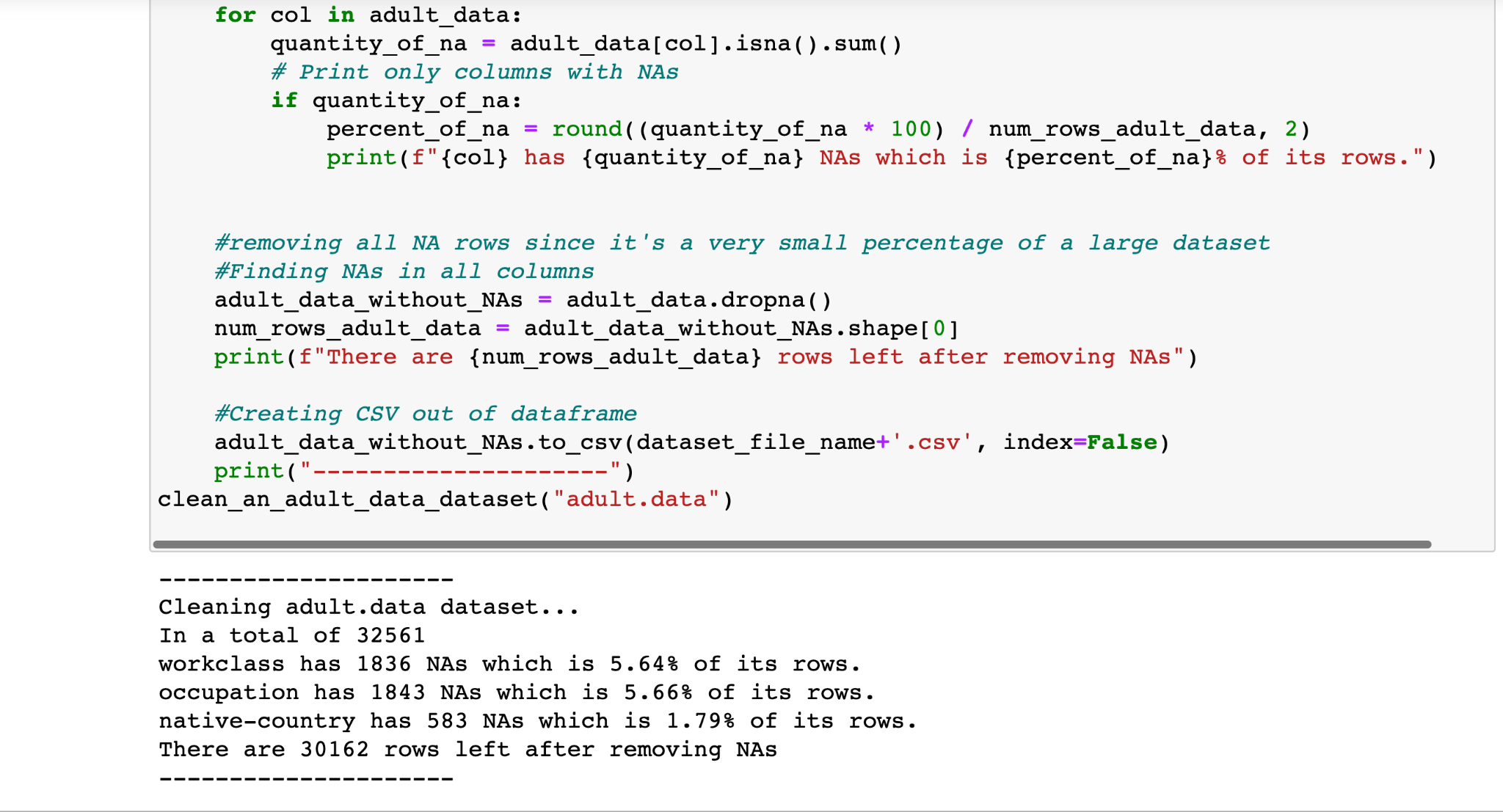
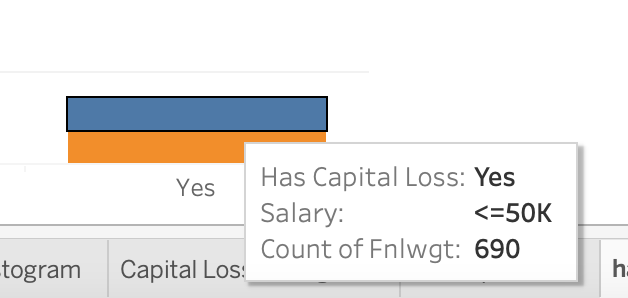
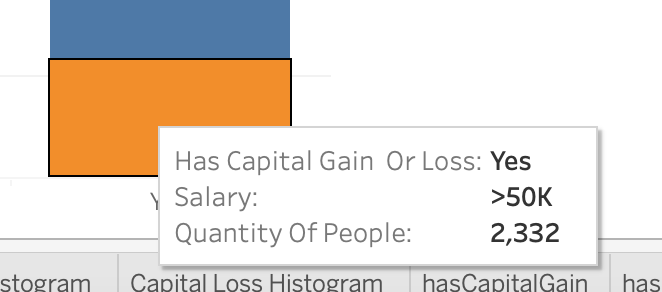


Image 1: Print of dataCleaning.ipynb execution

1. I started with an overview, realised that fnlwgt can be used as id since I put a count of it and it gave me the same number of registers as the total I have in the dataset on the Data Source Tab (Image 2).
2. Right after creating a chart to show how many people make more or less than 50k (Image 4), a bar chart sounded too simplistic and didn’t give proportions, just numbers. I’ve added an annotation saying the proportions and that was sufficient to make this more understandable for me. Edited the y axis to explain it means the quantity of people as CNT(Fntwgt) said nothing to me. The same guidelines where applied to Salary vs. capital gain/loss too.
3. Charts shown on the explain data helped me to realise things like everyone that makes near to 100k capital gain makes more than 50k salary a year, I tried to incorporate it on my final exploratory dashboard (Image 15) but it didn’t fit quite well but in any case this insight helped me to tell the final story on the explanatory dashboard.
4. On Quantity of people with Capital Gain on each Salary Group (Image 10), I’ve decided to create a categorical variable representing people who have and who don’t have Capital Gain, after plotting the plain quantities, I’ve noticed that it would be nice to add the salary group as an extra dimension. Adding it resulted in a colourless hard to understand chart. I’ve put colour differentiation on the salary groups to make this 3rd dimension more visible. I went back to the other charts with salary and added the same colour scheme to keep the consistency
5. I did a few charts like Capital Gain on each Salary Group (Image 10) without realising the tooltip was referring to “Count of Fnlwgt” which is not so readable, when I noticed it, I’ve changed the tooltip to say to say “Quantity of People” on this and other charts



## 

****

1. I’ve realised after making all my exploratory charts that I probably should have an extra chart showing that no individual has both capital gain and loss (Image 3) to assert that there is no data inconsistency (since a case like that would be impossible in the real world).
2. I’ve realised that I can edit the aliases of categorical variables like the Race one that had truncated names for Races like Amer-Indian-Esk which I’ve transformed into American Indian Eskimo and that helped in Image 13 and 14 to not make the reader thing “which race is this one exactly?”
3. I had to remove the annotations to make my charts be readable in the Exploratory Dashboard, unfortunately I couldn’t find a way to show them conditionally, only when the chart is not exposed on a dashboard for example. Now the annotations only exist in the prints that I put on this report.
4. For the Explanatory charts used on the dashboard in Image 16, I’ve decided to create a copy of the calculated field Has Capital Gain Or Loss and call it “Has Sold Assets” as it was a better name to use to tell the story of a possible relationship between wealth and income inequality.
5. I’ve decided to do brand new charts (Image 16) for the explanatory dashboard after trying with the old ones and not being able to tell a story, fitting annotations into it and putting colour on more enticing titles helped me to put something together.