P3 - Data Visualization in Tableau

o Data Visualization Fundamentals

- Data Vizualization Intro:
Data visualization is super awesome and important thing to understand large sets of data in the easiest way possible.

- Data types:
Before starting to create visualization, it's important to understand the types of data we work with. There are two types of data, quantitative and qualitative.

→ Quantitative:
Quantitative data are things you measure as numbers such as temprature, money, and no. of scratches from your cat. You can split quantitative data into two groups, continuous and discrete. Discrete data can take on only certain values. For example, the no. of items sold in a transaction can only be +ve whole number, you can't buy 2.5 pillows or 2.5 cats.

Continuous data can take on any value within some range, like time, height or money.

→ Qualitative:

Qualitative data is descriptive information about things that can't be quantified with numbers, like male/female and hair color. These are categorical data, data that indicates belonging to a category or group. Often you will want to group your data by the categories and compare them.

You also have ordinal data, things like rankings and subjectives scales you'd see in surveys, such as "How do you feel about tacos?"

1. gross!
2. ewww
3. okay
4. good
5. delicious

You will sometimes see ordinal data numbered 1 to 5, but the numbers don't really mean anything.

These are the typical data types we'll see. We could be working with things like images or text which might not seem to fit in these categories. But most of the time, but most of the time you can convert these into numbers or categories.

→ Summary of data types:

- Quantitative
  - Continuous
    - time, height, weight, money, interest rates temperatures
  - Discrete
    - unit sold, no. of languages spoken, no. of emails you received yesterday.

- Ordinal
  - Ranking, Survey questions like "How do you feel about cats?"
    1. Hate them
    2. Negative
    3. Neutral
    4. Positive
    5. Love them with all my being.

→ Flow Chart to identify data type

Do you have data?

Yes — No — Get Data

Is it numbers? — Yes

Words? — No — Yes

Is it a subjective Scale or ranking? — Yes — Qualitative, categorical data

No

Qualitative, ordinal data

Can the no.'s take on any value? — No — Quantitative, discrete data

Yes

Quantitative, Continous data

• Visual Encodings:
When we visualize our data, we use different type of visualization methods like:

- Dots
- Lines
- Bars
- Color, etc.

• Exploration Vs. Explanation

→ The Data Analysis Process
Visualizing data comes in two points during the analysis process: where you are exploring the data, and where you are explaining the data. Exploring involves digging through the data to find interesting relationship and questions. Explaining is when you present those relationships and answers to the questions. Typical process of working with data looks something like this:

Extract → Clean → Explore → Analyze → Share.

→ Extract and Clean
The first thing you do is extract the data from a database or paste it from records elsewhere. This is typically where we use things like SQL or collecting data from webpages, a process called Web scrapping.

Most of our time will be spent cleaning this data. Often, records will be missing, formatted

wrong, or just don't make any sense.

→ Exploratory Visualizations
Once the data is in good shape, you'll explore it to gain an understanding of the data. Let'll want to look at how data is distributed, if some variables are correlated, and how records are split between categories. This process is usually called EDA for exploratory data analysis.

Data Visualization comes in handy here as we can plot our distribution on our data, and create things like scatterplot to reveal correlations. It'll help us look for interesting patterns in the data, and other things that will help guide decisions.

→ Analyze
We use different plots to analyze our data like
- Bar graphs
- Histograms
- Scatter plots
- etc.

→ Explanatory Visualization.
The final part is where you look deeper into patterns you found during the EDA process and share them with people. This is explainatory part.

Try to tell the story from the data you analyzed to our audiences

- Design Principles
- Basic Figures

→ Geospatial Plots

Geospatial data (country, state, latitude & longitude) is readily viewed on a map. There are generally two types of maps, Choropleths use color to encode another value associated with the location on the map such as population, population density, GDP, etc.

Cartograms are similar to coropleths, but distort the boundary of regions - such as countries - to encode a value, typically along with some color encoding.

→ Small Multiples

A small multiple is a series of plots with the same scale that make it simple to compare data across groups. The plots can be practically anything: lines, bars, scatter plots, maps.

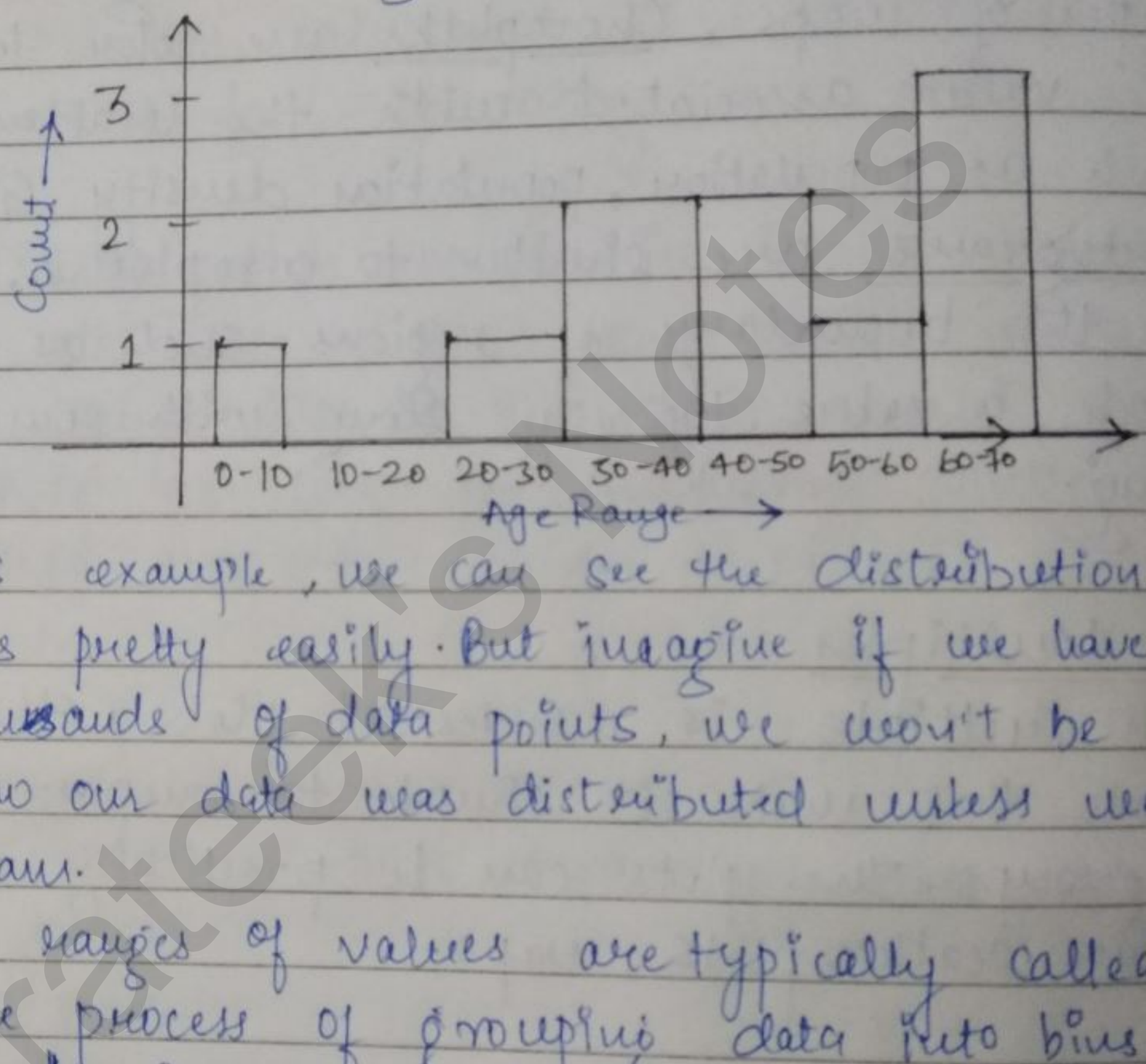
- Visualizing Distributions

Sometimes, showing the actual distribution of our data is optimal. Our distribution might not be normal! There could be outliers that strongly bias the mean. Simple things like bar charts can hide these issues.

→ Histograms

Histograms are bar charts built by grouping data into value ranges.

Ex: We have a group of people with ages: 29, 64, 44, 69, 31, 43, 32, 62, 8, 53. We can group the ages into 10 year ranges, then count up no. of people in each range.



In this example, we can see the distribution of the ages pretty easily. But imagine if we have hundreds or thousands of data points, we won't be able to tell how our data was distributed unless we use a histograms.

The ranges of values are typically called bins and the process of grouping data into bins is called binning.

We'll use histograms frequently to visualize distributions of continuous variables. The bin width and the placement of the bin edges can drastically effect how the distribution looks.

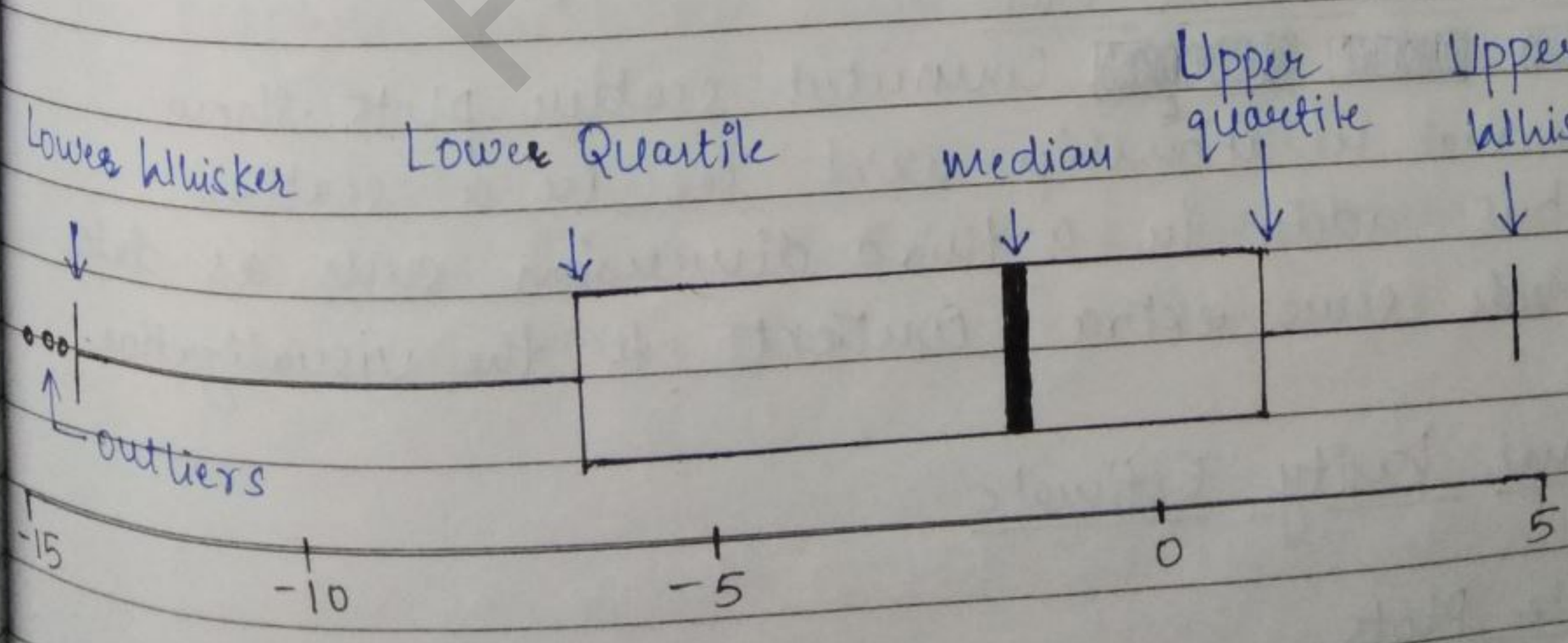
Binning has a nice property where it averages the data in the bins which can reduce noise, but if the bin width is too large, we might miss the fine details in the distribution.

However, if the bins are too narrow, there could be too much noise and interesting details might be lost. Also, the placing of the bin edges effect how the histogram looks. Sometimes we have to do the trial and error to get the bins right.

→ Box Plots

Box plots are common visualization for displaying the general shape of distribution using intervals. An interval is a value that is greater than some %age of the data. For instance, the 50% interval is the value greater than 50% of the data, typically called the median. The 95% interval is the value greater than 95% of the data.

All box plots use the 25%, 50%, 75% intervals, typically called quantiles. Usually, there will also be whiskers (or fences) that indicates some larger intervals, or the min and max. We'll often see box plots that show outliers, data points greater than or less than the whisker values.



o Other Awesome Graphs

There are few other graphs which are worth reading about, they are:

→ Bullet Graphs:

Developed by Stephen Few, bullet graphs are an extension of bar graphs that layers multiple measures on top of each other for comparison.

→ Sparkline:

Edward Tufte introduced these in visual display of Quantitative Information (along with naming small multiples) as a way of succinctly visualize some quantity changing over time.

Sparklines are simple and designed to emphasize the change in quantity in a small visual area. They have caught on and are widely used in finance. We'll often see these associated with stocks to display the stock price over time.

→ Connected Scatter Plot

Connected scatter plots show the normal relationship we'd see in a scatter plot, but adds in a third dimension such as date to provide some extra context to the visualization.

→ Kernal Desity Estimate

→ Cycle Plots

o Using Color

Choosing the color for our graphs is very important for making graph easy to understand.

→ The most commonly used color palette in python and MATLAB used to be jet palette.

→ Jet palette is not a good palette instead we should chose Linear palette which make data visualization on graphs much better to understand.

→ Few other good palettes are:
- Sequential Palettes
- Diverging Palettes
- Palettes for qualitative data.

o Design Integrity

→ Lie Factor

The lie factor compares the size of the effect in the graphic with the size of the effect in the data.

$$\text{Lie factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$