Info visual note

Week 1

Stakeholder -> collect/import/ingest -> tidy -> transform -> visual -> model -> communicate

Distribution of data : shape of data

Visualization stat: how data variables/features relate to each other

Distribution of data

* 1. An early step in any effort to analyze or model data should be to understand how the variables **are distributed**
  2. Techniques for distribution visualization can provide quick answers to many important questions:
     1. What range do the observation cover?
     2. What is their central tendency?
     3. Are the observation is heavily skewing in one direction?
        1. Symmetric: skew = 0, excess kirtosis
     4. Is there evidence of bi-modality?
     5. Is there any significant outliners?

Uni hist

Bio hist

Kde

ECDF

Box plot

Visualizing statistical relationships:

* 1. Stat analysis is a process of how variable in the dataset relate each other and how those relationship depend on other variables; p-value; B1 statistically != 0
  2. Visualize can be pro component can be process because when properly visualize can see trend and pattern indicated relationship
  3. Typically use two common approaches:
     1. Scatter plot
     2. Line plot

The empirical rule: m + 1sd = 65% + 2sd = 95% + 3sd = 99.7%

Trim the data to deal with outliner

(Done)Fundamentals of Data Visualization A Primer on Making Informative and Compelling Figures by Claus O. Wilke

**Amounts**: bars/dots; group/stack the bars if multiple categories

**Distribution:** histograms, density (arbitrary parameter choices); Boxplot, violin plots, strip chars, and sina plots for many distribution. Stacked hist and overlapping densities for smaller number of distributions. Ridgeline plots for large numbers

**Proportions:** pie charts, side-by-side bars, stacked bars; pies charts, grouped bars(moderate), stacked bars(large), stacked densities(continuous) for multiple variable. Mosaic plot, treemap, parallel sets.

**X-Y relationships**: scatterplots, density contours (large number) 2D Bins, Hex Bins, Correlogram, Line graph, connected scatterplot, smooth line graph.

**Geospatial Data**: map, choropleth, cartogram, cartogram heatmap

**Uncertainty**: Error bars, 2D Error Bars, graded error bars. Confidence strips, eyes, half-eyes, quantile dot plot, confidence band, graded confidence band, fitted draws.

Data Visualization Charts, Maps, and Interactive Graphics by Robert Grant 2018.pdf

1.1 because our brains are wired that way: patterns and difference

1.2 to help the analyst avoid problems: spotting anomalies

1.3 to win over the audience:

Why visualization of data:

**to give an overview,**

**to show the scale and complexity of the data,**

**to allow exploration of the data,**

**to communicate findings,**

**to tell a story,**

**to attract attention and stimulate interest**

Narrow down the candidates:

How many different observations and variables do I need to show?

Am I going to show the individual data, some aggregate of them, or some statistic?

Are the variables nominal, ordinal or scale

Do I want readers to focus on some comparison or change in the data

What formats are my reader likely to be familiar with

Work on visualization:

What is the message

What parts of the data are evidence for it

What other parts need to be shown for contrast/context?

Do I/we know how to do this, or can we learn it / adapt someone else's work, or do we need hire in

1.4 working on data visualization

The individuals who take on the job will come from either a statistical or a design background; think about which of these you need.

The people you hire can only do what you tell them you need, so clearly and comprehensively specifying the require- ments is essential.

Expect to be involved in user-testing, to look at early drafts and provide feedback.

Do not underestimate the tasks of design or web develop- ment.

Find out what the data are like before you start.

Get the timescale, budget, specifications and any confiden- tiality requirements agreed in writing first.

Propose a program of user-testing and feedback from the outset, and get agreement to meeting times for that.

Throughout, remember that your relationship to your clients is a professional one: they want you do a great job and to tell them what would be a good or bad idea.

1.5 A toolbox

* 1. Programmable statistical software like Stata or a data- focused programming language like R or Python,
  2. quick visualization software like Tableau,
  3. SVG editing in the text editor as well as a graphical interface   
     like Inkscape or Illustrator,
  4. sketching by hand,
  5. relevant JavaScript libraries – currently, D3 and Leaflet would be the choice,
  6. an online mapping tool, like Mapbox,
  7. a versatile program for big data and fast data, like Spark.

1.6 Be prepared to sketch and discard

Seaborn is a data visualization library built on top of Matplotlib. It is often used because it makes attractive visualizations and works easily with Pandas. While in Matplotlib you often had to write multiple lines of code to create a plot Seaborn makes assumptions on what you want which often translates into getting the same plot with 1 line of code.

**Distribution Plot**

# A univeriate distribution provides a distribution for one variable. Kernal Density Estimation with a Histogram is provided. Bins define how many buckets to divide the data up into between intervals

sns.distplot(crash\_df['not\_distracted'], kde=False, bins=25)

**Joint Plot**

# Joint plot compares 2 distributions and plots a scatter plot by default. With kind you can create a regression line with kind='reg'. 2D KDE with kind='kde'. a hexagon distribution with kind='hex'

**KDE Plot**

# Get just the KDE plot Kernal Density Estimation estimates the distribution of data

**Pair Plots**

# Pair Plot plots relationships across the entire data frames numerical values. With hue you can pass in a categorical column and the charts will be colorized

**Rug Plots**

# Plots a single column of datapoints in an array as sticks on an axis. With a rug plot you'll see a more dense number of lines where the amount is most common.

**Categorical Plots**

**Bar plots**

# Focus on distributions using categorical data in reference to one of the numerical columns

**Count plot**

# A count plot is like a bar plot, but the estimator is counting the number of occurances

**Box plot**

# A box plot allows you to compare different variables.

**Violin plot**

# Violin Plot is a combination of the boxplot and KDE While a box plot corresponds to data points, the violin plot uses the KDE estimation of the data points. Split allows you to compare how the categories compare to each other

**Strip Plot**

# The strip plot draws a scatter plot representing all data points where one variable is categorical. It is often used to show all observations with a box plot that represents the average distribution.

**Swarm plot**

# A swarm plot is like a strip plot, but points are adjusted so they don't overlap. It looks like a combination of the violin and strip plots

**Matrix Plots**

**Heatmaps**

plt.figure(figsize=(8,6))

sns.set\_context('paper', font\_scale=1.4)

# To create a heatmap with data you must have data set up as a matrix where variables

# are on the columns and rows

crash\_mx = crash\_df.corr()

**Cluster Map**

# A Cluster map is a hierarchically clustered heatmap. The distance between points is calculated, the closest are joined, and this continues for the next closest (It compares columns / rows of the heatmap)

Week 1:

Sub2\_BostonDF = BostonDF[["CRIM","INDUS","CHAS","TAX"]]

Sub2\_BostonDF.shape

Sub2\_BostonDF.describe().T

Sub2\_BostonDF.corr()

Week 2:

What is the KDE Plot?

1. The Kernel Distribution Estimation Plot (KDE) depicts the probability density function of a continuous variable . We can plot univariate KDEs or bivariate KDEs.

2. The Y-axis in a KDE plot represents the PDE (probability density estimate), defined as probability per unit value of whatever variable is on the X-axis.

What is it used for?

1. KDE Plots are used for visualizing the Probability Density of a continuous variable. A KDE depicts the probability density at different values in a continuous variable.

When is it used?

1. A kernel density estimate (KDE) plot is a method for visualizing the distribution of observations in a dataset, analagous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions.

2. The motivation behind the creation of KDEs was that Histograms are not smooth, they depend on the width of the bins and the endpoints of the bins, KDMs reduce the problem by providing smoother curves.This can be useful if you want to visualize just the “shape” of some data, as a kind of continuous replacement for the discrete histogram

3. KDEs are more flexible than Histograms. The smoothness of the kernel density estimate (compared to the discreteness of the histogram) illustrates how kernel density estimates converge faster to the true density for continuous random variables.

#Question 6: write the code to draw a shaded bi-variate KDE plot for weight and acceleration features in the cars dataframe

#Put the acceleration feature on the vertical axis and the weight feature on the horizontal axis

#Break out the KDE plot by the cylinders feature

#include a color car to additional provide insight on densities

sns.kdeplot(x=cars.weight, y=cars.acceleration, hue=cars.cylinders, shade = True, cbar = True)

#Question 10: Write the code to plotting a univariate Displot for the displacement feature in the cars dataframe

#Make the Displot a kernel density estimate plot instead of a histogram

#Breakout the Displot by the origin feature

#Facet the data by origin

sns.displot(x=cars.displacement, hue=cars.origin, col=cars.origin, kind='kde')

Kde plots:

* 1. Kernel: a kernel is a function that helps estimate an unknown probability density function
  2. Kde kernel options include: (1) gaussian (2) triangular (3) cosine (4) exponential
  3. As long as you have enough data, your choice of kernel really does not matter
  4. Bivariate kde plots are based on 2 independent random variables. The bi-variate kde plot reveals an apparent relationship between 2 variables
  5. The bi-variate kde actually has a 3-dimensional bell-shaped appearance - although presented in a 2-dimension format in seaborn
  6. Seaborn uses level curves or **contour lines** to display 3-d kde plots in 2-d format
  7. Contour lines show density: when you combine contours with colors on a Seaborn kde plot, you might get a better sense of density for both variables. By default, darker colored areas represent higher density values
  8. When you also shade (i.e. Shade = True in Seaborn), the darker the shade, the higher the density
  9. Bw, bw\_adjust, bw\_method arguments in seaborn
     1. Bw: a number bw is a curve smoothing parameter. It has been deprecated since seaborn v. 0.11.0 use bw\_adjust or bw\_method
     2. Bw\_adjust: a factor that multiplicatively scales the value chosen using bw\_method. Increasing bw\_adjust will make the curve smoother and vice versa
  10. The bandwidth or standard deviation of the smoothing kernel can produce a distorted representation of your data. The rule of thumb that sets the default bandwidth works best when the five distribution is smooth, unimodal and roughly bell-shaped.

Endogeneity

Visualize data

->

Correlation analysis => "Y and X are correlated"

->

Regression analysis => "There is a statistically significant relationship between X and Y "

->

Causation analysis => "X caused Y"

Exogeneity

Spurious regression

Corr(x,y) = cov(x,y) / @x@y

(control group, treatment group)

Science => experimental science -> e.g. Pharmacology

=>Observational science -> e.g. data science

Week 3:

Box plot

Boxplots are used to generate a summary of statistics for features

Boxplots are famous for helping visualize the 5-number summary for a feature

Boxplots also help us quickly check for the presence of outliers for a particular feature/variable

Violin Plots

Violin plots are used to visualize the distribution of quantitative variables or features in Python Seaborn.

You can think of a violin plot as a combination of a box plot and two KDE plots

The violin plot, like the box plot is a categorical distribution plot

Random control trials (RCT)

Science - experimental (medicinal, chemistry)

- observational (data science)

Train test split

DID difference in difference, reduce endogeneity

Visualizations - 1. visualize individual variables

2. Visualize statistical relationships between variables

Model: correlation, regression ,causation

* 1. Journal of Finance
  2. Journal of Financial economics
  3. Review of financial studies
  4. Journal of quantitative analysis

All models are wrong but some are useful - George E. P. Box

All visualizations are misleading without sufficient context

cars\_odd = cars[cars.cylinders.isin([1,3,5,7,9,11])]

sns.violinplot(x = cars\_odd.cylinders, y = cars\_odd.displacement, hue = cars\_odd.origin), plt.legend(loc = 2)

sns.boxplot(y='mpg', x = 'origin', hue = 'cylinders', data = cars\_odd)

Week 4

What is the swarm plot?

1.A swarm plot is a version of the scatter plot used for representing categorical values

2. Swarm plots are also called beeswarm plots, and are similar to strip plots in that they plot all of the data points of your numerical variables. Swarmplots tend to avoid plotting overlapping points

3. Unlike strip plots (which we discuss in Day 3), however, swarm plots attempt to avoid obscuring points by computing non-overlapping positions instead of adding random jitter. Swarm plots share many of the same advantages of strip plots, but without as much clutter to hide their salient features.

4. Spreading out the points in a non-overlapping fashion however, limits the number of points that can be plotted—there is only so much space on any particulare page. Additionally, the algorithm that calculates the positions is computationally expensive and so scales poorly as the number of points increases - so its advisable to use swarm plots when you have small samples

5. We can use the seaborn.swarmplot() to create swarm plots.

6. It is advisable to use the swarm plot when you are trying to avoid overplotting or points overlapping

7. It is not advisable to use this type of graph when the sample size is large

sns.swarmplot(x='origin', y='horsepower', hue='cylinders',

data=cars\_filtered, dodge=True

)

1. Strip plots have the disadvantage of not dealing with overplotting effectively - relative to swarm plots

2. We can use jitter and adjust the alpha (transparency) for strip plots in attempting to deal with overplotting - so helping us see the plots better

sns.stripplot(x=cars\_filtered.origin, y=cars\_filtered.weight, hue=cars\_filtered.cylinders,

dodge=True)

Week 5:

sns.scatterplot(x=CAS\_filtered.income, y=CAS\_filtered.expenditure, hue=CAS\_filtered.county)

sns.lineplot(x=Marijuana\_data.year, y=Marijuana\_data.checks, hue=Marijuana\_data.colour)

#Regression plots help you build and visualize a linear regression

#model for your data

#The whole point of a regplot is to overlay your scatter plot with

# a linear regression model

sns.regplot(x=Sample\_d.carat, y=Sample\_d.price, ci=None)

#Polynomial regression - you can plot a polynomial

#regression if you notice a curved relationship between

#relationship like we see between price and carat

#Plot a polynomial regression plot rather than a linear one

#Use the order argument to represent the power of the

#feature on the horizontal axis

#Order 1 : linear (default)

#Order 2+: polynomial

#Note the changes in the confidence intervals with change in order

sns.regplot(x=Sample\_d.carat, y=Sample\_d.price,

order=1)

#Increase size of dots and make them transparent since

#they overlap

#Change color of scatter plot to lightgray

sns.regplot(x=Sample\_d.carat, y=Sample\_d.price,order=3,

scatter\_kws= {'s': 100, 'alpha': 0.5, 'color': 'lightgray'})

**Course Schedule:**

Monday, Wednesday, Friday 1:25PM – 2:15PM

**Instructor:**

**Abel Iyasele**, PhD

Assistant Teaching Professor

Email: Abel.Iyasele@colorado.edu

**Office Hours:**

**Wednesdays, 10:30AM – 11:30AM**

**Room 139 of the Information Science Department (TLC Building)**

Policy on war

Select all that apply when it comes to visualizing amounts

Group of answer choices

**The most common approach is to use horizontal or vertical bars**

**Box plots can give a summary of the data but they do not show modality**

**Dots can also be placed at the location where corresponding bars would end**

Heatmaps must not be used

Violin plots cannot be used to represent all the information

**If there are two or more sets of categories for which we want to show amounts, we can group or stack the bars**

The most common approach to visualizing amounts (numerical values of categories) is using vertical or horizontal bars. TRUE/FALSE?

Group of answer choices

**True**

Which of these are true? Select all that apply.

Group of answer choices

**Histograms and density plots provide the most intuitive visualizations of a variable distribution**

Q-Q plots are always misleading

Histograms and density plots both require parameter choices

**If we require overall shifts among variable distributions, we can always use pie charts**

**Cumulative density plots can represent the data but may be difficult to interpret**

The violin plot and box plot, when  conditioned  on a categorical variable, do not allow us to visualizs of a variable at the same time. TRUE/FALSE

**False**

We can use the kernel density estimate (kde) plot to refine insights gained from the histogram of the same variable. TRUE/FALSE?

Group of answer choices

**True**

Which of these is true?

Group of answer choices

Violin plots combine the histogram and the boxplot

Boxplots are not useful when we want to visualize many distributions at once

For violin plots, the hue argument in Seaborn is used to condition the plot on the first categorical variable

**Stacked histograms and overlapping densities allow a more in-depth comparison of a smaller number of variable distributions**

The pairplot always gives perfect information about the relationships between two variables

Why should we employ data visualization to communicate with stakeholders? Select all that apply.

Group of answer choices

**To be taken seriously by stakeholders, in addition to presenting the data or insights to support communication, we need to show them the data to support our message**

**Human brains always suppress numerical data and look for visuals**

**The human brain can spot tiny clues about speed and distance, and can distill that information from what the eyes see almost instantaneously**

**Human brains are wired for seeing patterns and differences and also for understanding spatial relationships from patterns and differences**

**Data visualization can be compact, accessible to stakeholders**

**It is always best to restrict messages to numbers**

**Visualizations always employ complex tools to communicate findings**

Visualizations are useful in the early stages of statistical analysis. It is good to generate many charts for the benefit of your analysis and stakeholders. TRUE/FALSE?

Group of answer choices

False

Anscombe's quartet is an example of plots which show a dataset with exactly the same statistics but which mask great differences. In a pairplot, we may see a different picture from calculating the correlation between two variables. TRUE/FALSE?

Group of answer choices

True

Group of answer choices

**To give an overview**

**To tell a story**

**To disprove the simplicity of numbers**

**To communicate findings**

**To attract attention and stimulate interest**

**To demonstrate expertise of the presenter**

**To allow exploration of the data**

**To show the scale and complexity of the data**

The default kernel for Seaborn kde plot is Gaussian. TRUE/FALSE?

Group of answer choices

True

Kde kernels cannot be normal

**The kernel for a kde plot is a function that helps estimate a known probability density function**

**Data visualization involves visualizing distributions of variables and possible statistical relationships between variables**

Visualizing statistical relationships use only one common approach: scatterplots

**Kernel density estimate plots are used to show what the coree distribution of numerical variables are**

**Statistical analysis is a process of understanding how variables in a dataset relate to each other and how those relationships depend on other variables**

**Kernel density estimate (kde plot) kernel options include Gaussian, triangular, cosine, exponential**

**The Shade=True option in Seaborn provides a way to shade the area underneath the kernel density estimate plot**

**Bivariate kernel density estimate plots reveal the joint probability distribution for two random variables**

The kernel for a Kde plot helps estimate an unknown probability density function

Infographics are NOT data visualizations because they do not represent numbers in any way except writing them out in text. TRUE/FALSE?

Group of answer choices

True

**False**

Why is it important to sometimes layout information from simple to complex when presenting information?

Group of answer choices

It allows stakeholders to drill down as far as they want

Which of the following are good questions to ask yourself as you begin work on a visualization? Select all that apply.

Group of answer choices

**What other parts need to be shown for contrast/context?**

What stakeholder representatives need to be in the room?

**What is the message is the chart providing?**

**What is the context or overall environment for the visualization?**

**Do I/we know how to do this, or can we learn from it/adapt someone else's work or do we need to hire in?**

**What stakeholder questions am I trying to answer?**

How many stakeholders do I have?

Is this chart complex enough?

**What parts of the data are evidence for the message?**

Which of these show us an example of why data visualizations are important?

Group of answer choices

Gelman and Unwin (2012) did not give sufficient reasons why statisticians would want to create visualizations of data

Statisticians have no way to communicate their findings without charts

Frank Anscombe's artificial datasets, which all have the same means, standard deviations, and correlations would have been misleading without the production of charts for each of the datasets

Cumulative density plots are always easy to interpret relative to histograms and density plots

**Human brains are not very good at spotting anomalies**

For two or more sets of categories for which we want to show amounts, we can map the categories on the horizontal and vertical axes and show amounts by color. TRUE/FALSE?

Group of answer choices

True

Why can histograms and density plots be misleading? Select all that apply.

Group of answer choices

**They can be unfriendly to the human eye**

**They are not easy to interpret without color**

They do not require arbitrary parameters

They are not friendly to the human eye

**They do not allow an in-depth comparison of distributions across categories**

They can be hard to present

**They can require arbitrary parameters**

The rule that most normally distributed variables follow is called? Select all that apply.

Group of answer choices

The central dispersion rule

**The Gaussian estimate**

**The 68-95-99.7 percent rule**

The box and whisker rule

The proportion rule

The kde plot estimate

**The empirical rule**

The law of small numbers

What effective options can a statistician use to visualize amounts? Select all that apply.

Group of answer choices

Stacked correlogram chart

**Side-by-side bar charts**

Pie charts

**Stacked bar charts**

Stacked density charts

None of these answers

Correction:

Visualizations are useful in the early stages of statistical analysis. It is good to generate many charts for the benefit of your analysis and stakeholders. TRUE/FALSE?

True

Infographics are NOT data visualizations because they do not represent numbers in any way except writing them out in text. TRUE/FALSE?

True

Firms captial:

Capital = b0 + b1 firm

Delta capital = b2 + b3 delta firm

Capital t - capitcalt-1 = b2 + b3(firm st - firm st-1)

Week 6, Day 3 - Heatmaps:

Heatmaps - correlogram

(1) Variant of heatmaps that use the intersection of the vertical and horizontal axies to represent a possible linear relationship such as a linear correlation

(2) You can use correlogram in exploratory data analysis EDA to build descriptive or predictive models

* Clustered/clustering heatmaps (you can use to show density)

(1) The goal here is to build associations between both the data points and their features

(2) We want to see which features are similar or different from each other

#WHAT IS IT?:

#A heatmap is a way to visualize a table of numbers

#A heatmap is a way to vizualize data that presents magnitude of features as color in two dimensions.

#The variation in color in a heatmap may be by hue or intensity,

which gives obvious visual cues to a reader about how the feature is clustered or varies over space

#A heatmap is a representation of data in the form of a map or diagram in which data values are represented as colors

#An image or map representing the varying features

#A heatmap depicts values for a main variable of interest across two axis variables as a grid of colored squares.

The axis variables are divided into ranges like a bar chart or histogram,

#and each cell’s color indicates the value of the main variable in the corresponding cell range

#WHAT IS IT USED FOR?:

#To show how features are clustered or the associations/ relationships between two variables/features

#Used for visualizing a correlation matrix

#WHEN IS IT USED?:

#A heat map helps you visualize density. So use them when you want to show how data is clustered

#Heatmaps are used to show relationships between two variables, one plotted on each axis.

#By observing how cell colors change across each axis, you can observe if there are any patterns in value for one or both variables

sns.heatmap(Medical\_data.corr(),cmap='coolwarm',center=0, vmin=-1, vmax=1, annot=True)

Structure of the seaborn pairplot

Gives a relationship plot between 2 attributes/features/variables

Diagonal charts -> gives a probability distribution plot

Week 7 day 3:

2 key tasks in statistics and data science:

|  |  |
| --- | --- |
| Describe data | Aka descriptive statistics |

Infer about population from a sample. If infer about population parameters from sample statistics

|  |  |
| --- | --- |
| Generalize sample results to populations | Aka inferential statistics |

Population draw sample infer and generalize but using CLT

Central limit theorem CLT:

Population variable: normally distributed => sampling distribution of sample means will be normally distributed

Not normally distributed => sampling distribution of sample means will be normally distributed if (1) sample size n > 30 large enough (2) sample has finite variance

Prediction of CLT: m = m, theta = theta / sqrt n

#WHAT IS IT?:

1. The Pairplot is a tool to help visualize data to find any apparent relationships between specific variables where the variables are continuous or categorical. Pairplot visualization is useful when you need to do Exploratory data analysis (“EDA”).

2. The Seaborn pairplot combines various plots on one chart. You may see a scatterplot, a regplot, kde, histograms

#WHAT IS IT USED FOR?:

1. Plot pairwise relationships in a data-set

#WHEN IS IT USED?:

When you need a quick first-pass at the apparent relationship between continuous or categorical variables

sns.pairplot(Usedcar\_data, vars=['Age', 'Price', 'Mileage'], kind='reg', diag\_kind='kde')

sns.pairplot(tips,

kind='reg',

plot\_kws={'color':'xkcd:hot pink'},

diag\_kws={'color':'xkcd:beige'})

sns.pairplot(tips,

hue='weekend',

y\_vars=['tip'],

x\_vars=['total\_bill', 'size'],

kind='reg',

height=4)

Week 8 day 1

#WHAT IS IT?:

1. The joint plot combines the apparent relationship between two variables and the individual distributions per variable in one picture

2. The jointplot consists of two plots: (1) the relationship plot which gives a sense of the joint distribution of the two variables and (2) the per-variable histogram or distribution plot, which gives a sense of the marginal distributions of the individual variables

#WHAT IS IT USED FOR?:

1. Plotting joint and marginal distributions in one joint picture

#WHEN IS IT USED?:

1. When you need to see the marginal distribution for the variables in a dataset

2. When you need to see the joint distribution for pairs of variables in a dataset

Barplot

#WHAT IS IT?:

1. Seaborn barplot is the categorical estimate plot (CEP)

Week 9 Facet grid

Parent population:

Normal -> sample size irrelevant for CLT to hold

Non-normal -> sample size at least 30 for CLT to hold

-> sample must have finite variance for CLT to hold

#WHAT IS IT?:

1. Seaborn's FacetGrid serves as the backbone for the following plots:

(1) catplot

(2) relplot

(3) displot

2. Seaborn uses the concept of small multiples to build out Facet Grids that in turn suppport various plots like scatterplots etc

#WHAT IS IT USED FOR?:

1. Facet Grid is used to create a grid, on which we can map other types of plots with minimal code

#WHEN IS IT USED?:

1. When you need to build out various plots from scratch

I want to standardize a phrase I have used in the latest homework and for other work going forward.

Please note that when you see plot Limit-Income relationship or B-A relationship, it means plot Limit or B on the vertical axis and Income or A on the horizontal axis.

I know we have used it the other way around in past Jupyter Notebooks, but lets keep it the way its described above for the current Homework, Exam 2 and everything else we will cover till the end of the course.

Week 10 Cat plot: categorical plot

A facet grid - supported plot that helps plot a numeric - categorical relationship

Scatterploe: x, y are always numeric

Catplot: x is a categorical variable, y is usually numeric

#What is it?:

1. Seaborn's Cat plot stands for Categorical Plot that helps you build various kinds of categorical variable plots
2. With the Cat plot, you get acess to all of Seaborn's categorical plots including:

-boxplot -violinplot -stripplot -barplot -swarmplot

1. The Cat plot is also based on Seaborn's Facet Grid - which means you will be able to use some of the features of Seabon's Facet Grid

#When can you use Seaborn's Cat plot?

You can use it to create multiple categorical plots within Seaborn and change those plots in real time to better present your data

#What is It?

#1.Relplot stands for Relational Plot

#2. It is built on the Seaborn relational plot Facet Grid

#3. You largely have two options: (1) scatter plot and (2) line plot

#When do you use it?

#1. When you want more flexibility to build out or customize your scatterplot and line plots

# using Seaborn Facet Grid

Week 11: LmPlot and count plot

2 tailed 2 sided test: H0: M1 = M2 H2: M1 != M2

1-sided 1 tailed H0: M1 = M2, H1: M1 > M2, H2: M1 < M2

Y = b0 + b1x

Statisital significance x, sign/direction, magnitude

Hypothesis testing

Prediction

Description

#What is it?:

1. LM plot stands for linear model plot
2. The LM plot is basically a combination of the regression plot (regplot) and Facet Grid
3. The LM plot allows you to create Small Multiples, using the Facet Grid

#When can you use Seaborn's LM plot?

1. You use the LM plot when you want to visually represent the possible relationship between two numerical variables
2. You can also use the LM plot to visually show a logistic regression
3. The LM plot also can help us visually detect the presence of Simpson's Paradox ahead of any modeling activity

**Week 12, Day 1 - Clustermap**

#WHAT IS IT?: #A clustermap is a matrix plot that allow you visualize your matrix entities through a heatmap

You also get a clustering of your rows and columns in addition to a heatmap.

#WHAT IS IT USED FOR?:

The Seaborn cluster map is a matrix graphic that allows you to visualize your matrix elements as a heat map while simultaneously displaying a clustering of your rows and columns.

#WHEN IS IT USED?: Used to evaluate whether samples within a group are clustered together

Pairgrid

#WHAT IS IT?:

#A PairGrid is a matrix plot that allow you visualize your matrix entities through a heatmap

# You also get a clustering of your rows and columns in addition to a heatmap.

#WHAT IS IT USED FOR?:

#PairGrid allows us to draw a grid of subplots using the same plot type to visualize data.

# Unlike FacetGrid, it uses different pair of variable for each subplot.

#It forms a matrix of sub-plots. It is also sometimes called as “scatterplot matrix”

#WHEN IS IT USED?:

#The Seaborn PairGrid allows us to plot pairwise relationships between variables within a dataset with greater flexibility

# that the pairplot