DATA VISUALIZATION USING TABLEAU

ASSIGNMENT 1

SEPTEMBER 10, 2022.

INDIANA UNIVERSITY
(BLOOMINGTON)

Authored by: VARSHA R
IU ID: 2000751388
Sem: Fall 2022

Introduction

Data visualization is achieved using Tableau, a software package focusing on business intelligence (BI). The word tableau refers to a graphic representation or description. As a novice with Tableau, I have experimented with the software and built a few visualizations. The purpose of this report is to discuss these visualizations. For each graph, the following information will be discussed:

- Understanding of the dataset.
- Plots built using these datasets.
- Plot analysis.
- Inferences drawn from the visualizations.

Dataset

> The UCI Machine Leaning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. - UCI

For the visualizations, we will use the following datasets from the UCI Machine Learning Repository:

- Census: A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). Prediction task is to determine whether a person makes over 50K a year.

- Mushroom: This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one.

- Iris: This is perhaps the best-known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

- Car: The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes: buying, maint, doors, persons, lug_boot, safety. Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods.

Visualizations

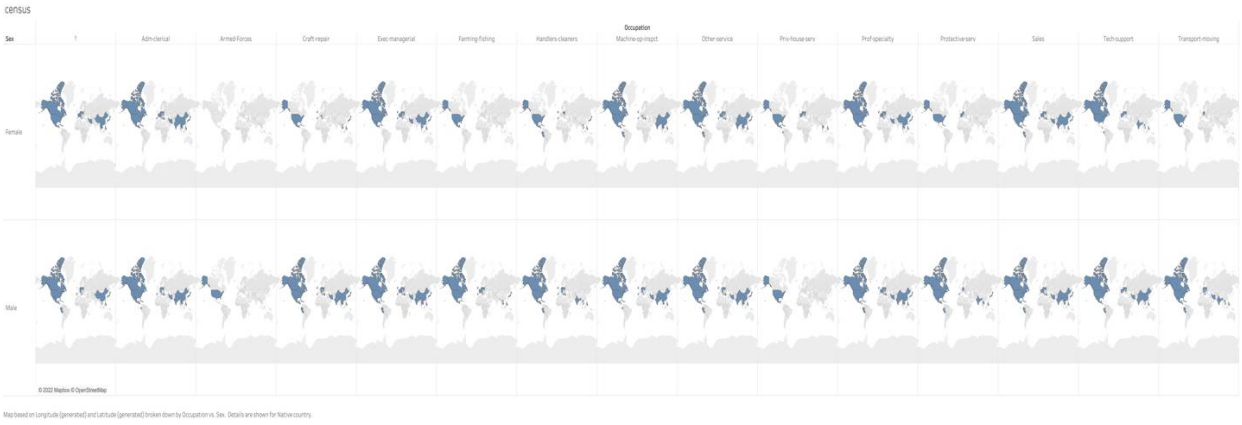
1) Census dataset

○ Understanding of the dataset

Dataset involves using personal details such as education level to predict whether an individual will earn more or less than $50,000 per year. There are 32,561 rows of data. You can find missing values that are marked with '?'. It contains the following columns:

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: 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: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: 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: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, etc.

○ Plot built using this dataset



Map based on Longitude (generated) and Latitude (generated) broken down by Occupation vs. Sex. Details are shown for Native country.

- ## Plot analysis

Attributes used for this plot: Native country, sex and occupation.
Columns: Occupation, generated longitude
Rows: Sex, generated latitude
Map: generated based on native country attribute data

In order to compare and visualize the ratio of males and females working different jobs across the world, this plot was visualized. A geographical type of graph is the most appropriate for this purpose, as it shows which countries have specific types of occupations and which genders contribute more to each occupation around the world. The graphs show that the occupation is primarily located in North America as compared to the other regions. Additionally, we can compare which occupations are dominated by males across the map, such as transport-moving, protective services, handlers-cleaners, etc. In addition, we can observe that there are no highlights in the female section of the armed forces graph. As you examine the graph further, you will notice that occupations are distributed differently across the maps for males and females.

- ## Inferences drawn from the visualizations

- Males work most of the physically challenging jobs across the world.
- There are no females working in armed forces.
- All of the males having armed forces as an occupation are from United States.
- United states is the only country in the dataset that has people working all the mentioned occupations.
- Female take the lead in private house service occupation worked across the globe.

2) Mushroom dataset

- ## Understanding of the dataset

Dataset involves details described in terms of physical characteristics. We can use these characteristics to determine if a mushroom is edible or not. There are 8124 rows of data. No missing values can be found in this dataset. It contains the following columns:

- cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
- cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
- cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
- bruises?: bruises=t,no=f
- odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
- gill-attachment: attached=a,descending=d,free=f,notched=n
- gill-spacing: close=c,crowded=w,distant=d
- gill-size: broad=b,narrow=n

- gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
- stalk-shape: enlarging=e,tapering=t
- stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
- stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
- stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
- veil-type: partial=p,universal=u
- veil-color: brown=n,orange=o,white=w,yellow=y
- ring-number: none=n,one=o,two=t
- ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
- spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
- population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
- habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Plot built using this dataset

mushroom

Habitat / Stalk-Shape

F1	d		g		l		m	p		u		w
	e	t	e	t	e	t	e	e	t	e	t	e
e	56	1,824	640	768	240		256	136		96		192
p	692	576	596	144	16	576	36	432	576	128	144	

Count of mushroom.csv

16 1,824

Count of mushroom.csv broken down by Habitat and Stalk-Shape vs. F1. Color shows count of mushroom.csv. The marks are labeled by count of mushroom.csv.

Plot analysis

Attributes used for this plot: Habitat, stalk shape and edible/poisonous.
Columns: Habitat, stalk shape
Rows: Edible/poisonous

Based on this graph, we can observe the concentration of mushrooms found in different habitats. Based on the attributes we are observing, we can also observe the blank spaces in the graph where no mushrooms are found. Additionally, the graph shows that urban and waste habitats have very few mushrooms, whereas the concentration increases across the left side. With respect to concentration, we can observe that woods and

grasses have the greatest number of mushrooms, and woods have the greatest number of edible mushrooms.

○ Inferences drawn from the visualizations

- In the woods, tapering-shaped mushrooms, specifically 1824 mushrooms, were found to be the most edible.
- The majority of mushrooms can be found in wooded areas and grassy areas, while the least amount can be found in waste areas.
- As far as edible mushrooms are concerned, there are no tapering stalk-shaped edible mushrooms found in leaves, meadows, paths, urban or waste habitats, while poisonous, enlarging stalk-shaped mushrooms can be found in waste.

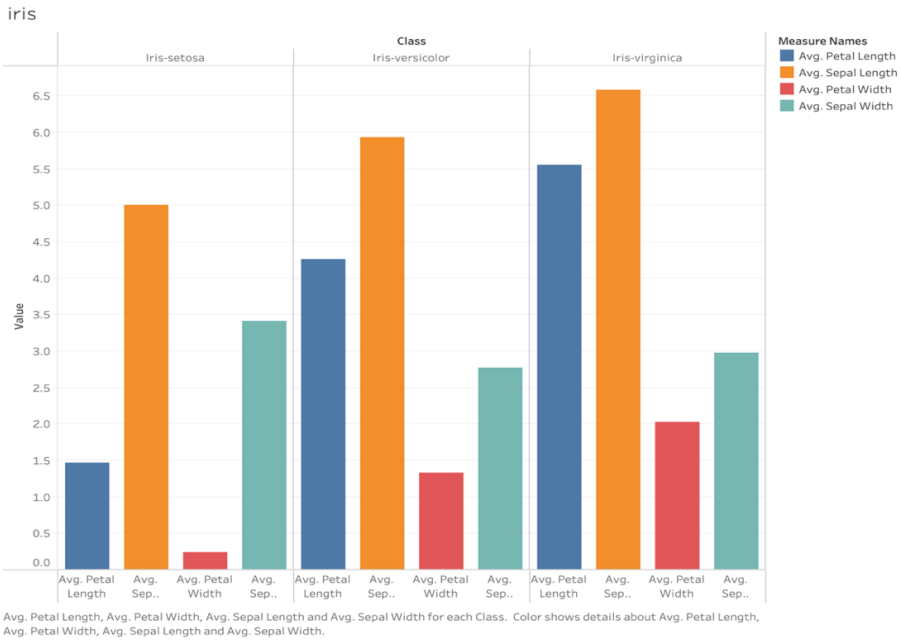
3) Iris dataset

○ Understanding of the dataset

Iris dataset is famous flower data set which was introduced in 1936. It is multivariate classification. There are 150 rows of data. No missing values can be found in this dataset. It contains the following columns:

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm
- class:Iris Setosa, Iris Versicolour and Iris Virginica

○ Plot built using this dataset



- **Plot analysis.**

  Attributes used for this plot: Class, width and length of sepal, width and length of petal
  Columns: Class, measure names
  Rows: Average measure values of width and length of sepal, width and length of petal

  This graph illustrates the average width and length of both sepals and petals for each class of iris. There is no doubt that sepal length dominates all classes of iris whereas petal width is the least dominant. The longest sepals are found in Virginica. As shown in the table below, we can summarize the graph as follows:

Class	Avg. Petal Le..	Avg. Sepal L..	Avg. Petal W..	Avg. Sepal W..
Iris-setosa	1.464	5.006	0.244	3.418
Iris-versicolor	4.260	5.936	1.326	2.770
Iris-virginica	5.552	6.588	2.026	2.974

Avg. Petal Length, Avg. Petal Width, Avg. Sepal Length and Avg. Sepal Width broken down by Class.

- **Inferences drawn from the visualizations.**

  - The petal width of Setosa is the smallest and the sepal width is the largest.
  - Comparatively, Versicolor has an average width and does not possess any dominant characteristics.
  - The Virginica has the longest sepals, the longest petals, and the widest petals.

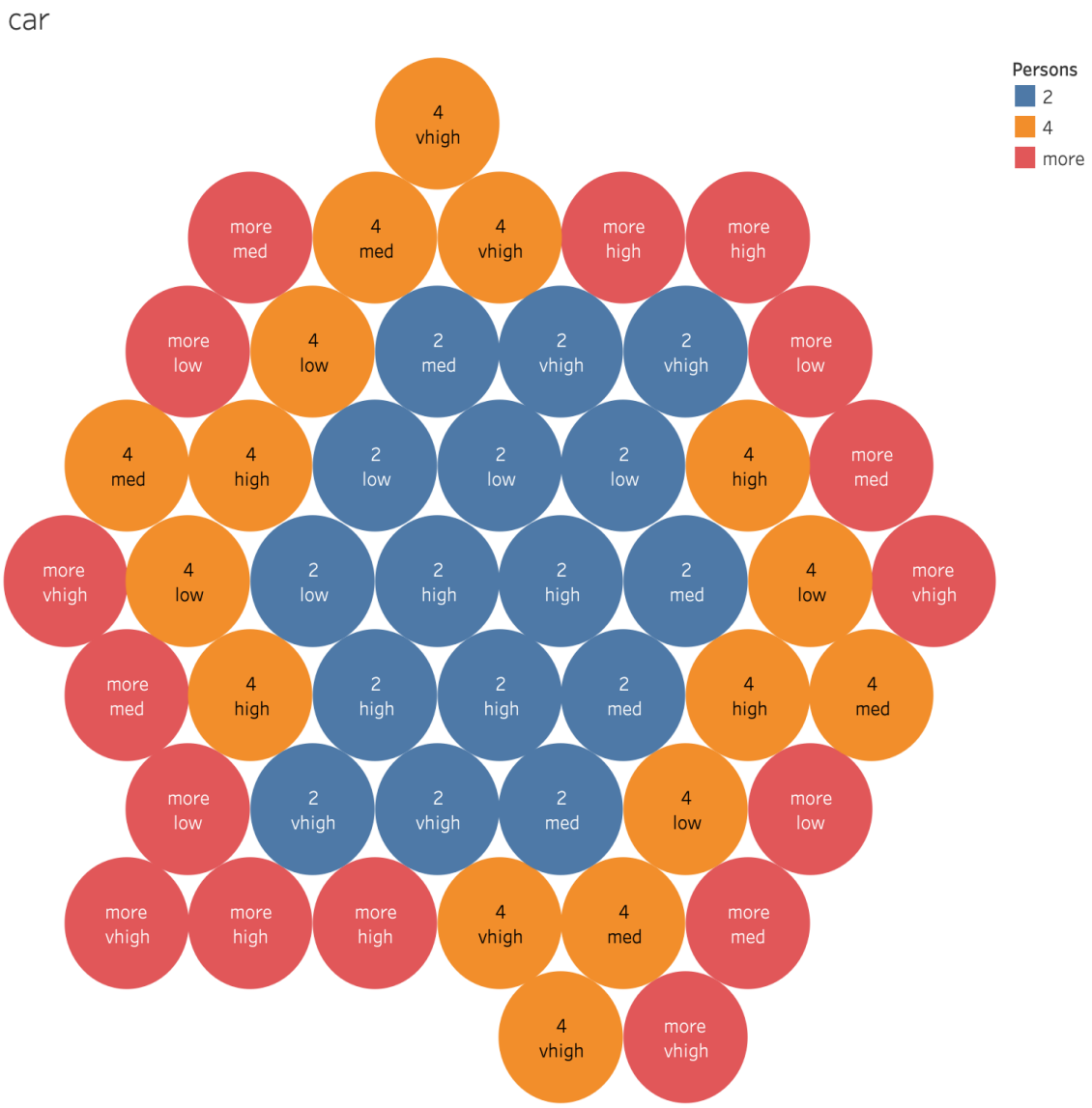
4) Car dataset

- **Understanding of the dataset**

  Inductive induction and structure discovery methods may be tested by utilizing this database derived from a simple hierarchical decision model. Data is presented in 1728 rows. In this dataset, there are no missing values. Columns include:

  - Class Values: unacc, acc, good, vgood
  - buying: vhigh, high, med, low.
  - maint: vhigh, high, med, low.
  - doors: 2, 3, 4, 5more.
  - persons: 2, 4, more.
  - lug_boot: small, med, big.
  - safety: low, med, high.

- **Plot built using this dataset**



Persons, Buying and Maint. Color shows details about Persons. Size shows count of car.csv. The marks are labeled by Persons, Buying and Maint.

- **Plot analysis.**

Attributes used for this plot: Person, buying and maintaining

My objective here was to compare and check how much the costs of buying and maintaining a car would be based on the number of people that can fit in the vehicle. There is a trend here in this dataset that, regardless of what attribute is used to divide the dataset, it always gives symmetrical values for the whole table/graph. The following table illustrates this:

Persons	Maint	Buying high	low	med	vhigh
2	high	36	36	36	36
	low	36	36	36	36
	med	36	36	36	36
	vhigh	36	36	36	36
4	high	36	36	36	36
	low	36	36	36	36
	med	36	36	36	36
	vhigh	36	36	36	36
more	high	36	36	36	36
	low	36	36	36	36
	med	36	36	36	36
	vhigh	36	36	36	36

Count of car.csv broken down by Buying vs. Persons and Maint.

- **Inferences drawn from the visualizations.**
  - There is a wide range of EQUAL car options on the market that can be tailored to meet your preferences.
  - Any visualization analysis performed on this dataset would result in the same distributed value across all attributes.

References

http://archive.ics.uci.edu/ml/index.php