Data Science with Python
Ticket Out the Door
Set 1: Summary Statistics Key

Name _____ Period _____

Skill 1.1 Exercise 1

Let's imagine we are working for the city government of the fictional city of Melody Metropolis. The mayor of Melody Metropolis wants to know more about the musicians who currently live in the city.

Open the dataset below and make a copy.

https://docs.google.com/spreadsheets/d/1m9XbRZ40Deds2HUm4S04hFPjUywZpRfZMII2TbXlFic/edit?usp=sharing

How would you describe this dataset? In your group, discuss the following,

- What does a typical musician's income look like? - ~ 20,000 – 50,000 seems to be the majority
- Is there a wide range of musician ages? - yes
- What proportion of the musicians in the dataset play guitar? - ~ 20%, but it is hard to say without further analysis

How accurate are your answers to the questions above?

Skill 1.2 Exercise 1

The variable instrument from our dataset tells us the type of instrument each musician plays. The table below gives the frequency, proportion, and percentage of musicians that play each type of instrument.

Fill in the missing entries to complete the table. Round all proportions to two decimal places and all percentages to the nearest whole number with the % symbol included after.

Instrument Categories

INSTRUMENT	FREQUENCY	PROPORTION	PERCENTAGE
voice	262	0.27	27%
guitar	278	.29	29%
piano	178	.19	19%
drums	135	0.14	14%
saxophone	48	0.05	5%
violin	57	0.06	6%
TOTAL	**958**	**1.00**	**100%**

What is the ratio of voice to violin?

262/57 ~ 5

What is the ratio of guitar to drums?

278/137 ~ 2

Data Science with Python
Ticket Out the Door
Set 1: Summary Statistics Key

Name _____ Period _____

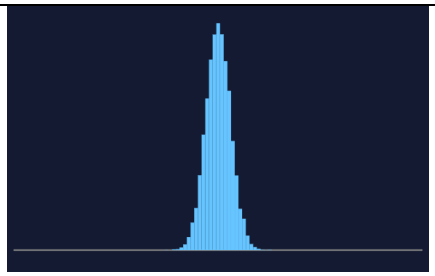
Skill 1.3 Exercise 1

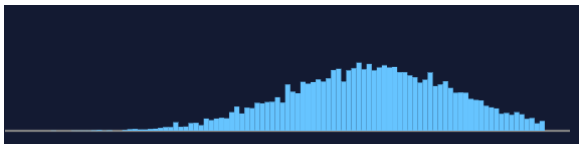
Refer to the musician dataset. Which variables are categorical? Which are numerical?

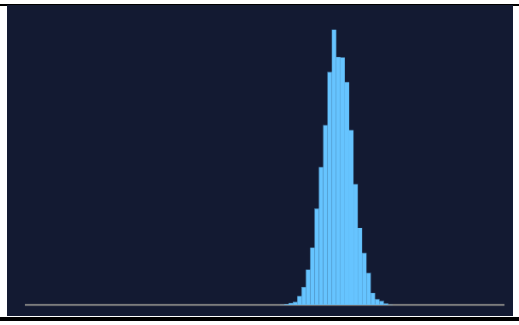
Categorical are non-numerical data -> title, instrument
Numerical – age, income, experience
*band could be either

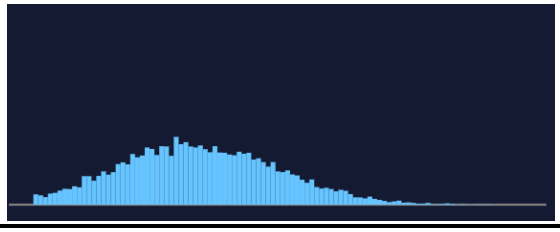
Skill 1.4 Exercise 1

Refer to the datasets below to answer the following,

A. 

B. 

C. 

D. 

For which dataset(s) is/are the value of the mean the highest?

B, C

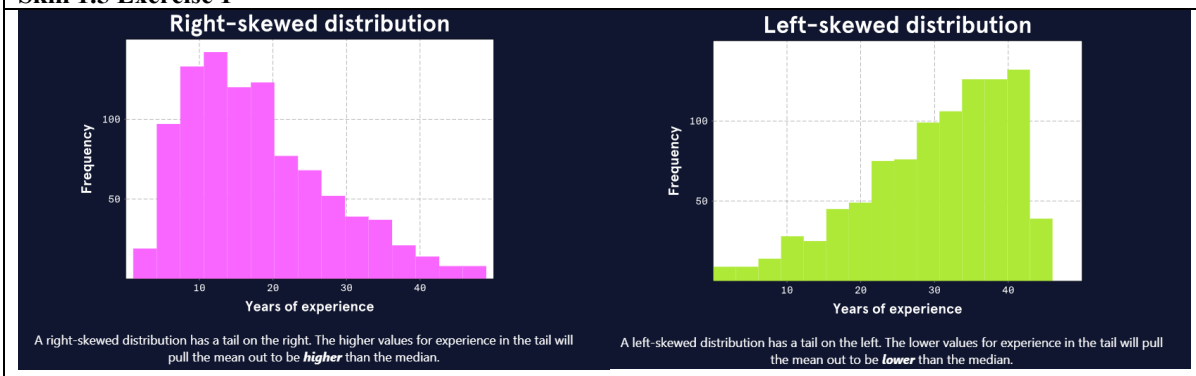
For which dataset(s) is/are the value of the standard deviation the greatest?

B, D

Data Science with Python
Ticket Out the Door
Set 1: Summary Statistics Key

Name _____ Period _____

Skill 1.5 Exercise 1



Another numeric variable from the musician dataset is years of experience working in the field of music. The plot on the left shows what the experience distribution might look like if it is right-skewed like the income distribution is. The plot on the right shows what the experience distribution might look like if it is left-skewed.

Refer to the musician dataset. Which distribution most likely is true for the musicians in Melody Metropolis?

Looks like its right skewed

Skill 1.6 Exercise 1

Return to the musician dataset and navigate to the *Experience* tab. Do the following,

- Highlight all the data
- From the data menu, select "sort range" → "A to Z"
- From the data menu, select "Column stats" and record the following,

Average	18.05
Median (Q2)	15.5
Median of all the values less than Q2 (Q1)	11
Median of all the values greater than Q2 (Q3)	26
Difference between Q1 and Q3 (IQR)	15

Are the years of experience right skewed or left skewed?

Right skewed

Data Science with Python
Ticket Out the Door
Set 1: Summary Statistics Key

Name _____ Period _____

Skill 1.7 Exercise 1

The second plot above shows that the median appears almost unaffected by the addition of these three gigantic incomes: the median moves from $32,978 to $33,011. However, the mean makes a drastic change from $34,795 to $228,235.

Which alternative statistics could your colleague use in this case?

The mean is influenced by extreme values because every value is included in its calculation. Statistics that are less influenced by extreme values rely on alternative calculations, like using the rank of the value rather than the value itself.

Using the rank of the value, or **rank transformation**, in statistical analysis involves replacing the original numerical or ordinal data with their corresponding ranks within the dataset. This means that the smallest value gets a rank of 1, the second smallest gets a rank of 2, and so on. If there are tied values (two or more observations share the same value), they are assigned the average of the ranks they would have occupied.
For example, if you have the data set (3.4, 5.1, 2.6, 7.3), the ranks would be (2, 3, 1, 4) because 2.6 is the smallest (rank 1), 3.4 is the second smallest (rank 2), and so on.

Skill 1.8 Exercise 1

Return to the musician dataset and navigate to the Experience and Instrument tab.

What is the average experience for each type of instrument?

Instrument	**Average**
Drums	17.8
Guitar	17.4
Piano	18.8
Saxophone	21.5
Violin	22.75
Voice	17.4

Data Science with Python
Ticket Out the Door
Set 1: Summary Statistics Key

Name _____ Period _____

Skill 1.9 Exercise 1
Return to the musician dataset and navigate to the Age and Income tab. Do the following, • Highlight all the data • From the Insert menu, select Chart
Is the correlation between these variables positive or negative? Is the correlation strong or weak? The correlation is about 0, indicating a weak correlation