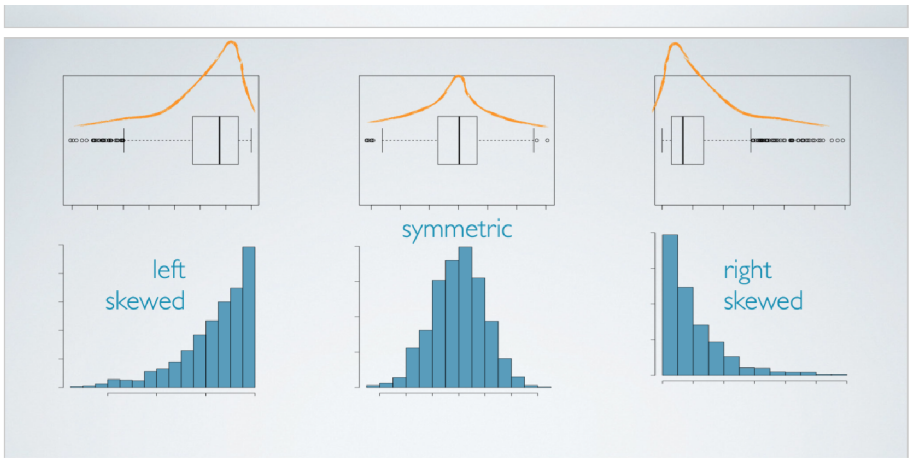
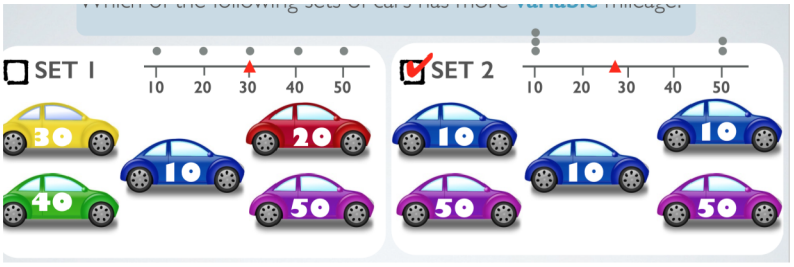
Course 1 - Week 2 - Visualization

Explore Numerical Variables

1. Describe scatter plots (motivation, how to use it)
    a. Scatter plots are used to explore the correlation (not causation!) of **numerical** paired data
    b. X axis is *explanatory variable* while Y axis is *response variable*
    c. To evaluate the relationship between the two variables, we draw a curve to fit the data pairs and then focus on the following four characteristics of the curve: **direction, shape, strength, outliers**
2. Describe histograms (motivation, characteristics, how to use it)
    a. Histograms are used to present the **data density** of **numerical** variables
    b. Skewness
        i. left skewed
            1. long tail on the left (data is “pulled” towards the left side)
            2. mean < median < peak
        ii. right skewed
            1. long tail on the right (data is “pulled” towards the right side)
            2. peak < median < mean
        iii. symmetric
            1. mean = median = peak
    c. Modality
        i. Number of peaks
        ii. Unimodal, bimodal, uniform, multimodal
    d. Bin width
        i. Too small: missing overall trend
        ii. Too large: missing important details
3. When do we use dotplot?
    a. It is useful when individual values are of interest
    b. Doesn’t work well when sample size increases
4. Describe box plots (motivation, characteristics, how to use it)
    a. It is useful for outlier detection
    b. We have median (50%) and Interquartile range (IQR, 25%, 75%) in the plot
    c. Illustration of box plots and histogram together



5. Why do we use square difference in the variance calculation?
    a. In pure difference calculation, positive difference and negative difference might cancel out
    b. Compared to absolute value difference, square difference gives more weights to large deviations
6. What's the difference between variability and diversity?
    a. Variability relates to the measure of centre and depends on how far each point is away from the centre (mean, mode, median)
    b. Diversity relates to how many different types of objects are there in the sample.
    c. In the figure below, set 1 has more diversity while set 2 has more variability

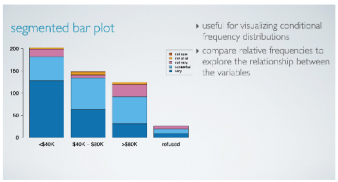


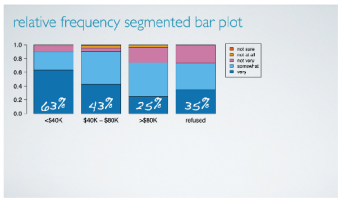
7. What is robust statistics?
    a. Measures (such as mean, variance, median, etc) on which extreme values have little effect
    b. Robust measures: median, IQR
    c. Non-robust measures: mean, SD, range
8. Why do we use data transformations?
    a. To see the data structure from a different way
    b. To reduce skewness (log transformation applies to data where a lot of points are close to 0)
    c. Straighten non-linear relationships to scatter plots

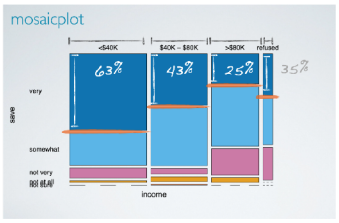
Explore Numerical Variables

9. What do we use to describe the distribution of a single categorical variable?
    a. Use bar plot (different from histogram!)

10. What do we use to evaluate the relationship between two categorical variable







11. What do we use to evaluate the relationship between a categorical variable and a numerical variable
    a. side-by-side boxplot