Data 100

Lecture 7: EDA & Visualization
Exploratory Data Analysis (EDA)

“Getting to know the data”

A process of transforming, visualizing, and summarizing data to:

- Build/confirm understanding of the data
- Identify and address potential issues in data
- Inform the subsequent analysis
- Discover potential relationships

EDA is an open-ended analysis
- Be willing to find something surprising
Exploratory Data Analysis (EDA)

“Getting to know the data”

- We used EDA with the CO₂ data and DAWN data to check the quality of the data.
- We also use EDA to help prepare for formal modeling.
- We also use EDA to confirm our modeling was reasonable.
- Plots can uncover features, distributions, and relationships that can’t be detected from numerical summaries.
John Tukey
Princeton Mathematician & Statistician

Introduced
- Fast Fourier Transform
- Exploratory Data Analysis
- “Bit”: binary digit

Early Data Scientist

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine
EDA is like detective work

“Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there.”

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine
EDA is Active and Incisive

“Exploratory data analysis is actively incisive rather than passively descriptive, with real emphasis on the discovery of the unexpected.”

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine
The Variable Represents
The Variable Represents

Urban Dictionary:

Go and be a good example to the others of your group or in your position

Huh?

A Variable represents a feature

It is distinct from it’s coding in a data file or data frame. It is more than a column in a table.
**Variable**

- **Quantitative**
  - Continuous: Could be measured to arbitrary precision.
    - Examples: Price, Temperature
  - Discrete: Finite possible values
    - Examples: Number of siblings, Years of education

- **Qualitative**
  - Ordinal: Categories w/ levels but no consistent meaning to difference
    - Examples: Preferences, Level of education
  - Nominal: Categories w/ no specific ordering.
    - Examples: Political Affiliation, CalD number

Note that categorical variable can have numeric levels and quantitative variables may be stored as strings.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Quantitative Continuous</th>
<th>Quantitative Discrete</th>
<th>Qualitative Nominal</th>
<th>Qualitative Ordinal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO\textsubscript{2} level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of siblings</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>GPA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income bracket</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td>X X</td>
</tr>
<tr>
<td>Number of years of education</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yelp Rating</td>
<td></td>
<td></td>
<td></td>
<td>X X</td>
</tr>
<tr>
<td>Lane of traffic (left, middle, right)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Basic Plots

Match Variable Type to Plot Type
Basic Visualizations

- How to choose the “right” one(s)
- How to read them –
  - Distributions
  - Relationships
Kaiser Study

- Oakland Kaiser mothers
- 1960s
- Measure the babies' weight (in ounces) at birth

All babies:
- Male
- Single births (no twins, etc.)
- Survived 28 days

Data provenance:
Mothers who use Kaiser
Starts as administrative dataset, expanded into study with new data
Selection mechanism **not** random
Information collected on mother’s and their babies

- Birth weight (ounces) **Quantitative Continuous**
- Gestation (weeks)
- Parity - total number of previous pregnancies **Quantitative Discrete**
- Mother’s height and weight
- Mother’s smoking status
- Mother’s age, race, education level, income level
- Father’s information and more… **Qualitative**

**Qualitative Ordinal**
One Variable

What is the Distribution of the values of the variable?
Quantitative – Continuous

- Birthweight
- The most basic visual representation of one quantitative variable is the rug plot

Hard to see much of the distribution with this rug plot

one thread for each observation
Birthweight

Histogram

With the histogram we hide the details of individual observations and view the general features of the distribution.

How would we describe the distribution of birth weight?
Distribution Features

- **Modes**
  - Number: 1
  - Location: near 120 oz
  - Size: main mode

- **Symmetry**
  - Symmetric
  - Slightly left or right

- **Tails**
  - Long, short, “normal”

- **Gaps**

- **Outliers**

![Normalized birth weight distribution of babies](image)
Distributions & Smoothing
A Small Dataset

10 values
0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8

Rug Plot
Shows the location of each value
We want to smooth these rug threads

BECAUSE

- this is a sample and we believe that other values near the ones we observed are reasonable
- we want to focus on general structure rather than individual observations
Important Properties of Histograms

- Total Area of the bars = 100% (or 1)
- Units on the y-axis are percent/x-unit
- Area of a bar = percentage of values in that bar unit matching:

\[ \text{Area} = \frac{x \text{ km wide} \times y \% / \text{km}}{\text{Area in } \%} \]

\[ = \frac{x \times y \%}{\text{Area in } \%} \]
Example - One large bin from 0 to 5

The 10 points are spread evenly across one large bin

\[ \text{Area} = 5 \text{ km} \times 0.2 \text{ fraction} \]

= 1 fraction

Not very informative distribution
Example

Bin width \( \frac{1}{4} \text{ km} \)

With these narrow bins, the histogram is little more than a rug plot

\[
\text{Area} = \frac{1}{4} \text{ km} \times 0.4 \text{ Frac/km} = 0.1 \text{ Frac of sample}
\]

\[
\text{Area} = \frac{1}{4} \times 0.8 = 0.2 \text{ (2 observations)}
\]
Example

Bins can be different widths

Area = 2 km \times 0.15/\text{km} = 0.3
A histogram smooths these points because:

- this is a sample and we believe that other values near the ones we observed are reasonable
- we want to focus on general structure rather than individual observations

The values 3.1, 3.6, and 4.8 have their proportion (3/10) spread over the bin [3,5]. That is, without the rug, we can’t tell where the points are in the bin.
Kernel Density Estimate: Alternative Smoother

Consider one point

Smooth with a kernel function, rather than in a histogram bin

Area under this curve is 1 or 100%
3 points – each represents 1/3 of the data

Place a kernel with area 1/3 on each point

Normalize the kernels by the number of observations
KDE – 3 points

Sum the 3 kernels at each point to get the density curve

\[ f(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) \]

- \( K_h \) is the green Kernel function
- \( h \) refers to how peaked/spread the kernel is
Example

Flat kernel

Density Curve is akin to 1 bar histogram

Broad Flat Density Curve

Kernel centered at
Example

Density Curve is too grainy. It's like a rug plot.

The software chooses a kernel bandwidth for you, but you can also specify your own.
Compare the Histogram and the KDE

0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8
How would we describe the distribution of birth weight?

Unimodal

Main mode at 120 oz

Slight left skew

Tails about normal
Histories and Density Curves

Describes distribution of data – relative prevalence of values

- **Histogram**
  - relative frequency of values
  - Tradeoff of bin sizes

- **Rug Plot**
  - Shows the actual data locations

- **Smoothed density estimator**
  - Tradeoff of “bandwidth” parameter (more on this later)
Box Plot

- Useful for summarizing distributions and comparing multiple distributions

Outliers are more than 1.5 * IQR away from lower and upper quartiles.

Visualization of summary statistics

Can lose a lot of features, such as...

- Modes
- Gaps
Our Data
0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8

- Median
- Lower Quartile
- IQR
- Hinge

Tukey's short cut for finding these values

\[ n = \text{#obs} = 10 \]

The median is the average of the \( \frac{n+1}{2} \) smallest or largest observations.

\[ \frac{10+1}{2} = 5.5 \]

Average the 5th and 6th:

\[ \text{median} = \frac{2.2 + 2.8}{2} = 2.5 \]
Our Data

0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8

- Median
- Lower Quartile
- IQR
- Hinge

IQR is $VQ - LQ = 3.1 - 0.9 = 2.2$

Hinge is $1.5 \times IQR = 1.5 \times 2.2 = 3.3$

To find the quartiles, we take the countdown value for the median, drop the $\frac{1}{2}$ (if it has it), and add one & divide by 2, e.g.,

$5.5 \rightarrow \frac{5 + 1}{2} = 3$

LQ is 3 in from bottom
UQ is 3 in from top

Any value more than 3.3 away from LQ/UQ is an outlier
Quartiles from Tukey’s “depth”

- Depth of the Median = \( \frac{n + 1}{2} \)
  - Count in from top or bottom of ordered set of values
  - If depth has a half then average the two values on either side

- Depth of Quartile = \( \frac{\text{round}(m) + 1}{2} \)
  - Round the median depth down to nearest integer
  - Count in from bottom to get the LQ and from the top to get the UQ
  - If depth has a half in it then average the two values on either side
Percentile – Need a more general def

- The $P^{th}$ percentile of a set of data is:

**Smallest** value that has **at least** $P\%$ of the data **at or below** it

0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8

10$^{th}\%$tile = 0.7  
90$^{th}\%$tile = 3.6  
60$^{th}\%$tile = 2.8

15$^{th}\%$tile = 0.8  
83$^{rd}\%$tile = 3.6  
66$^{th}\%$tile = 2.9

Notice the percentile will always correspond to a data point.

*Any value below 3.6 will not have 83$^{rd}\%$tile below it.*
Percentile – with weighted data

The P\textsuperscript{th} percentile of a set of data is:

**Smallest value** that has **at least** P% of the data at or below it

5. 5. 5. 5. 5. 5. 20. 20. 20. 20. 50. 50.

50\textsuperscript{th} percentile = 20

75\textsuperscript{th} percentile = 20

4/8 = 50\% \text{ is at or below 20}

3/8 = 37.5 \% \text{ is at or below 5} \Rightarrow \text{nothing in between}
Quantitative Discrete

We look for the same features

- Symmetry and skew
- Modes (number, location, and size)
- Tails (long, short, normal)
- Gaps
- Outliers
Discrete Quantitative  # of Siblings

What’s the difference between these 2 plots?

Bar plot – height not area is the proportion
Qualitative

We look at the relative size of groups

- Equally distributed
- Symmetry, Modes, Tails and Gaps don’t make sense
- Do most fall in one group?

Answers have implications in building prediction models
Qualitative Variable  

Bar Width – has no meaning

Education level

Dot plot focuses on comparison of the values

Why do we not reorder the bars according from shortest to tallest?
Pairs of Variables

Combinations:

Both qualitative,

One qualitative and one Quantitative,

Both Qualitative
Plotting Pairs of Quantitative Variables

- Scatter plot uncovers form of relationship between 2 variables
- Linear relationships are particularly simple to interpret
- Simple and elegant statistical theory for linear relationships
- Models are typically approximations, choose a simpler model over a complex one
Common Relationships

- **Ideal**
  - Can also have more spread

- **Simple Linear**

- **Simple Nonlinear**
  - Good too
  - Typically we transform to a linear rel.

- **Unequal Spread**
  - Still linear but we need to take care when modeling

- **Complex Nonlinear**
  - Very difficult to work with
The scatter plot is a 2-d rug plot.

Height measured to nearest inch so we get those stripes.
Hex Bin

Shading corresponds to density of points in the cell.

Why hexagons -
- Easier to see elliptical/linear relationships
- More efficient for covering region
- Visual bias of squares - drawn to see vertical and horizontal lines

Histogram of each variable
Smooth Contour

kde in 2-d

Contours of the 3-d density smooth of our 2-d data
Smoothing Scatter plots

Now we want to smooth the y-values as a function of x.
Smoothing Scatter plots

For an x-value consider all of the x’s near it.
Take an average of the y-values.
Smoothing Scatter plots

For an x-value consider all of the x’s near it
Take an average of the y-values
Smoothing Scatter plots
Create bins for all x
Average y-values in each bin
Smoothing Scatter plots

These averages sketch out a curve

Connected these blue
Smoothing Scatter plots

Rather than a simple average in fixed bins

We use kernels positioned on the $x_i$ to determine the weights to place on the $y_i$ in the average

$$g(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_h(x - x_i)y_i}{\sum K_h(x - x_i)}$$

The denominator ensures the weights sum to 1
Smoothing Scatter plots

Rather than a simple average

We use kernels positioned on the $x_i$ to determine the weights to place on the $y_i$ in the average

$$g(x) = \frac{\sum_{i=1}^{n} K_h(x-x_i)y_i}{\sum_{i=1}^{n} K_h(x-x_i)}$$

The denominator ensures the weights sum to 1
Smoothing Scatter plots

For each $x$, we find $g(x)$ by a weighted average of the $y_i$.

The $y_i$ are weighted according to the kernel function.

So $x_i$ far from $x$ do not contribute much to $g(x)$.

\[
g(x) = \frac{n}{\sum_{i=1}^{n} K_h(x-x_i)} \sum_{i=1}^{n} K_h(x-x_i)y_i\]
Local Smoothing

- Moving window
- Smooth/Average y values in the window
- Many different approaches for doing this:
  - kernel methods (what we just showed),
  - cubic splines, thin plate splines,
  - Locally weighted smooth scatterplot (lowess)

Allows us to see shape of the relationship between y and x
Mix Quantitative & Qualitative
Mix Quantitative & Qualitative

Side-by-side Boxplots

Side-by-side violin plots

Quantitative

Can't see modes

Can see tails and asymmetry

Can put Qualitative Variable on the y-axis

Density on its side
Mix of Qualitative and Quantitative

Overlaid bars/curves

Hard to tell one curve from another
Two Qualitative Variables
Pairs of Qualitative Variables

What’s the difference between these 2 plots?

Smoking status normalized within Education level

The plot on the right is normalized.
Interaction/Factor Plot

Smoking status normalized within Education level

Lines connect values for each group so easier to compare across categories on the x-axis.

Values in line to compare.

No width.
### Univariate Graphical Displays

<table>
<thead>
<tr>
<th>Type</th>
<th>Plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeric – few observations</td>
<td>few observations&lt;br&gt;Histogram, Density curve&lt;br&gt;Box plot, Violin plot&lt;br&gt;Normal quantile plot&lt;br&gt;Few Observations - Rug plot, Dot plot&lt;br&gt;Caution if discrete: density curves and box plots may be misleading</td>
</tr>
<tr>
<td>Categorical – Counts of categories</td>
<td>Dot chart&lt;br&gt;Bar chart&lt;br&gt;Pie chart (avoid!)&lt;br&gt;Caution if ordinal – order of bars, dots, etc. should reflect category order</td>
</tr>
</tbody>
</table>
# Bivariate Graphical Displays

<table>
<thead>
<tr>
<th>Numeric</th>
<th>Categorical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeric</td>
<td>Scatter plot</td>
</tr>
<tr>
<td></td>
<td>Smooth scatter</td>
</tr>
<tr>
<td></td>
<td>Contour plot</td>
</tr>
<tr>
<td></td>
<td>Smooth lines and curves</td>
</tr>
<tr>
<td>Categorical</td>
<td>Side-by-side bar plot</td>
</tr>
<tr>
<td></td>
<td>Overlaid Lines plot</td>
</tr>
<tr>
<td></td>
<td>Side-by-side dot chart</td>
</tr>
</tbody>
</table>
Caution about EDA

With enough data, if you look hard enough you will find something “interesting”

Important to differentiate inferential conclusions about world from exploratory analysis of data
Take care with EDA

- EDA can provide valuable insights about the data and data collection process

**BUT**

- Be cautious about drawing/reporting conclusions
  - Recognize that EDA biases your view
  - Be careful about sharing plots or hypothesis without additional validation ...

- Have a lot of data? Apply EDA to sample of the data before conducting formal analysis.