Data Storytelling

- Data Storytelling – The translation of complex data preparation and data analysis into a narrative that can be easily understood by the intended audience
- Allows us to communicate our findings
- Supports decision-making
- Evidence-based and descriptive
- Ideally, someone can replicate the results based on description
- Mediums of storytelling:
  - Visualization
  - Narrative

Visualization

- Visualization – Presentation of data in a graphical format.
- Ex. A graph

Narrative

- Narrative – The art of presenting data in an accessible and engaging manner
- A well-crafted narrative makes the information more accessible to a diverse audience
- Written and spoken
- Sometimes people will shape their presentation to fit their ideal narrative
- It is important for people ot be critical when presented with information

Big Data

- There is a lot of data around us everyday
  - 402.74 million terabytes of data is generated each day on the internet
  - Google manages 20+ petabytes of data each day
  - All of Wikipedia compressed is only 24.05 gigabytes
  - 750+ billion images on the internet
  - 214 million emails are sent on the internet every minute
- The vast amounts of data available can be difficult to manage
- Big data techniques assist with dealing with large amounts of data and processing data in real-time (streaming)
- This course won't focus too much on this, but be aware that there are techniques to handle large amounts of data

Data Science Around Us

Data Science in Society and Industry

- How data science could benefit people
  - Amazon reviews assist customers in making more informed decisions
  - Social scientists monitor videos featuring cultural trends found on social media
  - Impactful technological innovations from online forums provide business insights
  - Web-scraping supports the study of technological trends
  - Machine generated data influence public-sector administration

Data Science Close to You

- Fitbits (wearable technology) can collect data about our health
- Smartphones have GPS data
- Applications on smartphones can collect a large variety of data
- The collection of personal data brings up questions about privacy and ethical use of data

Data's Importance to Business and Industry

- Data shows marketers which product categories responded well to past disruptions and reveal which still need to be disrupted
- Data can show how best to engage their customers (contests, news, giveaways)
- Graphic designers use A/B testing to validate their designs
- Professional sellers may focus on predicting sales leads
- Data science skills can be applied to almost every job and academic field

Ethics

Risks in Data Science

- We will revisit ethics throughout the course
- Think about what information that might have ethical concerns in the following:
  - Email spam filters to divert unwanted emails
  - Medical information collected of patients
  - Web browser collecting all data about websites visited and search queries
- Data should be collected ethically
- Two general ethical concerns in data science:

  - Issue of privacy
  - Issues related to fairness / bias in decision-making algorithms
    * Machine learning models could be trained on data that contain biases or the model learns to predict based on cultural bias (Garbage in, garbage out)

Dataframes

- Dataframe – A tabular data structure with named columns (similar to a spreadsheet)
- Dataframes are fairly commonly used in data science and often provide additional functionality
- Other software / applications can be designed to provide functionality for analyzing and plotting through dataframes
- One common dataframe is through the pandas library

Pandas

- pandas is a library in Python that can use dataframes
- `import pandas as pd`
- `data = {'col_1': [10, 20], 'col_2': ['a', 'b']}`
- `my_dataframe = pd.DataFrame(data)`

Selecting Rows and Columns

- `data = {'col_1': [10, 20, 30], 'col_2': ['a', 'b', 'c']}`
- `my_df = pd.DataFrame(data)`
- `my_df['col_1'] # get values from column 1`

```
0    10.0
1    20.0
2    30.0
Name: col_1, dtype: float64
```

- `my_df.iloc[1] # get values from row index 1`

```
col_1    20.0
col_2    b
Name: 1, dtype: object
```

Functionality with Dataframes

- Mean of all values

- `my_df['col_1'].mean()`

  ```
  np.float64(20.0)
  ```

- `my_df.iloc[1].mean()`

  ```
  ERROR
  # can't calculate the mean of 20 and 'b'
  ```

- Let's try with new data
- `data = {'col_1': [10, 20, 30], 'col_2': ['a', 5, 'c']}`
- `my_df = pd.DataFrame(data)`
- `my_df.iloc[1].mean()`

  ```
  np.float64(12.5)
  ```

- Get unique values
- `data = {'col_1': [10, 20, 30, 10, 20], 'col_2': [4, 5, 'c', 1, 2]}`
- `my_df = pd.DataFrame(data)`
- `my_df['col_1'].unique()`

  ```
  array([10, 20, 30])
  ```

- Get the names of columns
- `my_df.columns`

  ```
  Index(['col_1', 'col_2'], dtype='object')
  ```