Lesson 6 - Applying the Data Science Method:

**Question for Mentor:**

* Data Wrangling: print percent of unique values per column – Is this asking me to print the percent of values that are truly unique aka the only one in that column?
* Stuck on appropriate fill value for fastEight column. Can’t figure out how to sum all columns of the chairlifts for the specified row
* Filling fastEight column – says there are fastEight lift values missing because the variance between total chairs and the rest of the chairs is zero, but big sky has a variance of 1.
  + Figured out what I did wrong – only 1 mountain with fastEight chair? I left it out of my sum that I subtracted from total chairs so it looked like there was a variance
* Are pairplots useful for datasets like this? Or should I be subsetting the columns?
  + Did I do this step right?
* Box plots don’t look too useful. Did I do it correctly?
* Were the clusters used for anything?
* Was able to complete step 4, but the Modeling steps were way over my head. Normal?
* Step 5 modeling, my numbers are out of control when I run the explained\_variance\_score and mean\_absolute\_error, also don’t know what they mean
* How do I verify that the numbers are right? I don’t have experience with these values so don’t know what even looks directionally right
* Says to get rid of base elevation, but that column was dropped in step 3 due to the correlation being >0.95

**Step One: Problem Identification:**

* Project done

**Step Two: Data Wrangling:**

* Four steps:
  1. Data Collection
     + Ask yourself, does that data I have allow me to answer the question of interest from problem identification?
     + Will want to load into friendly format like dataframe
     + Collating all data into single data frame for analysis will make project much easier going forward
     + Collating data early allows for clean data processing and easy adjustments later
  2. Data Organization
     + Directory organization is key especially with multiple iterations of same model
     + Keep things organized, clean and dated/versioned
     + Version control is a key concept in data science and should be mastered
  3. Data Definition
     + Defining data prior to development informs data science practitioner at a glance about their dataset
     + Data definition should contain following items by column:
       - Column name
       - Data type (numeric, categorical, timestamp, etc)
       - Description of column
       - Count or percent per unique values or codes (including NA)
       - Range of values or codes
  4. Data Cleaning
     + Most common types of data cleaning are:
       - Handling missing and NA data
       - Removing duplicates
     + Reviewing percentage of mission observations aids in determining the best step forward
       - Review percentage of observations missing per column
       - Drop, impute, or replace missing values
     + Drop duplicate row if they are repeated measures of the same observation

**Step 3: Exploratory Data Analysis:**

* Exploratory data analysis workflow can be split into 3 main groups
  + Build data profile tables and plots
  + Explore data relationships
    - Can be found using covariance matrices
  + Identification and creation of features
    - The goal of identifying features is to use your exploratory work to isolate features that will be most helpful in constructing a predictive model
* Exploratory data analysis (EDA) is often an iterative process where you pose a question, review the data, and develop further questions to investigate before beginning model development work

**Eploratory Data Analysis in Python:**

* Replace method replaces specific values with what we want
  + Syntax: pounds.replace([98, 99], np.nan)
* Inplace=True means you can replace existing series without doing pounds = pounds.replace([98, 99], np.nan)
* Boolean series
  + True = 1
  + False = 0
* You can select from a pandas dataframe using a Boolean array using .values
  + Syntax: full\_term\_weight = birth\_weight[full\_term.values]
    - This gives the birth weight of all full term babies
* Probability Mass Functions
  + Probability you get a value x
* Pmf class shows possible values in a distribution and their probabilities
  + Syntax: Pmf(variable, other arguments (i.e. normalize=true etc) )
* Cumulative distribution functions
  + Probability you get a value <= x
* Cdf class takes a variable and returns a CDF object
* Can take inverse by doing .inverse()
  + Syntax: variable.inverse(0.75) to get the 75th percentile
* CDFs are smoother than PMFs – takes out randomness
* CDFs give clearer view of distribution and are good for comparing distributions
* PDF = probability density function
* KDE = kernel density estimator
* Use CDFs for exploration
* Use PMFs if there are small number of unique values
* Use KDE if there are a lot of values

**Guided Capstone Step 3:**

* Histogram observations
  + Highly correlated: base elev & summit elev, runs & total chairs, days open and pass prices?
  + Duplicated: maybe base & summit elevation
* K-means clustering
  + Clustering essentially finds patterns in data when we don't know in advance what we're looking for. The K-means algorithm is one way of doing clustering
  + Will be covered in detail in the Unsupervised learning portion of the class

**Step 4: Pre-processing and training data development:**

* Pre-processing is the concept of standardizing your model development dataset
* 3 steps:
  + Create dummy or indicator features for categorical variables
    - Categorical values are those with data type ‘object’
  + Standardize the magnitude of numeric features
    - Standardize the numeric features so those with greater magnitude don’t take over the analysis
    - How to scale features?
      * Standardisation – replacing values by their z scores (# of standard deviations from the mean)
      * Mean normalization – distance from mean / range (max – min)
      * Min-Max scaling – distance from min / range
      * Unit vector – x / ||x||
    - When to scale
      * Any algorithm computing distance or assumming normality
  + Split into testing and training datasets
    - Split into testing and training datasets using a logical subset of data

**Guided Capstone Step 4:**

* Training Data
  + Need to learn more here

**Step 5: Modeling:**

* If your response variable is continuous, e.g. temperature or sales price, a numeric value that has a distribution closer to Gaussian than Bernoulli, than you will apply a regression type of model; otherwise a supervised classification model should be used
* A screenshot of a cell phone

  Description automatically generated

**Step 6: Documentation:**

* Reviewing modeling results
* Presenting and sharing your findings (data storytelling)
* Finalizing code
* Finalizing model documentation
* Great presentations shape



* Switches between what is and what could be until the end where you describe the new bliss – utopia with the idea adopted