# Prep:

Keep data zip file and zip file of videos handy – in case someone can’t access Canvas

# Day 1: Course Welcome

* 0900-0930
* Put on Screen ice breaker
  + Name, Org, One Data-related story (from work, media, research)
* This is course revamp
  + Motivation: My background on starting DS projects
  + Course times and days
    - We may not cover it all
* Go over Syllabus
* Ask if everyone has access to Canvas and a computer with software
* Go over website – how to play videos (show link)

Module 1: Lesson 1 and 2

* 0930-1030
* Intro Slides
  + Show AI.gov website
* DS Workflow slides
  + Give examples of places you works on the **Data Driven Missions** slide
  + Give examples of bad examples of Stakeholder needs (“here’s some data, see what you can find”)
  + Remind them of the DS workflow before you play audio
    - Ask them to write down when they see specific workflow stages in the video
      * Warn them that you’ll be asking for these after the video
  + Remember to **turn on Computer Audio** to play the DAU video
    - Offer the video link in case video doesn’t come through right
  + Have discussion after the video on what stages the students identified in the video
    - Give out “curiosity points” for people that answer
* SMART Questions slides
  + These are questions you can ask during the stakeholder meeting to refine your data science goal.
    - And these are questions to ask throughout your project as well.
* 10 minutes – Quiz and Break

Module 2: Lesson 1 (1045-1115)

* Return from break, Go over Module 1 Quiz
* M2L1: Installation and Setup
  + Ask who was able to install Anaconda
    - If everyone has it, then go over the following quickly
    - If not, then allow them time in the middle to get things installed
  + Show installation instructions (show link)
  + Go to Anaconda website à Download
    - Show “getting started guide” and “video introduction”
    - Show how to open Jupyter NB once it’s installed
    - Execute **code, markdown**,
    - Show script on how to reset default startup folder
      * Anaconda Command line: **jupyter notebook --notebook-dir="C:\Users\New Folder"**
    - Show how to install packages
  + Show Anaconda online notebooks (new feature)
    - Examples of what you uploaded
    - Need an account
    - Works with library file
  + Show Kaggle.com
    - Data files and challenges
      * Let’s you see how other data scientists work
    - Create account
    - Open Notebook
  + Also VAULT is possible on NIPR computers
    - Allows CUI and SIPR data
    - Post slides on how to create an account (One LSX onedrive file)
    - Not free as far as I know
  + Library File
  + Curiosity Points

Module 2: Lesson 2

Numpy Fundamentals (1115-1145)

* Explain black hole picture – Very fast and good for large volumes of data.
  + Link on bottom of page
* Go over Numpy slides
* Go over NP lab with students
  + Call on students to return matrix values
* Remember: Axis = 0 looks vertically (max for each column)
* Show them Shift+Tab to see description of any function
* Show Numpy Cheat Sheet

Pandas Fundamentals (1145-1215)

* Go over M2L2 Numpy and Pandas lessons
* Remember: iloc and numpy indexing doesn’t return the last value
  + Titanic.iloc[0:3,’age’] returns age for rows 0,1,and 2 only (not 3)
* Show Cheat Sheet

Advanced Pandas (1215-1300):

* Have Kaggle California Housing Prices ready: <https://www.kaggle.com/datasets/camnugent/california-housing-prices>
* Show cheat sheets and links on bottom of page

# Day2

M2L3 – (0900-0930)

* Go over word search

M2L3 – Matplotlib (0930-1015)

* Go over Plotting Slides
* Show cheat sheet in Jupyter Notebook, and show <https://matplotlib.org>

M2L3 – Seaborn (1015-1100)

* Show seaborn website <https://seaborn.pydata.org/>
* Ask why there is a line of house prices at the top of the scatter plot of **median\_income** vs **median\_house\_value**
  + Because house prices are capped at $500K. There are many houses worth more than that, but their true value is not known.

M2L3 – Advanced Plotting + break (1100-1115)

* Go to Plotly weblink
  + It is FREE
  + Show examples + code on bottom of examples
  + Don’t have them do the lab – keep it optional
* Module 2 Quiz

M3L1 – Prepare and Plan (1115-1130)

* Play starting video <https://youtu.be/ixIoDYVfKA0?si=EoieQMCFDSc41hfx>
* Go over slides

M3L2 – Process (1130-1230)

* Go over Slides (15 min)
* Missing data – if > 30% data, then probably consider dropping
  + Cabin is 70% missing, and you can get information indirectly from other variables like Fair (higher fair = first class cabin)
  + Age is 17% missing but it’s pretty useful so can’t ignore this variable.
    - Find ways to fill in missing values
    - Median no good because ages went from 6 months to 80 years
    - (Last slide) some researchers divided passengers by salutation and then processed the missing values.
* Labs: Let them watch and follow
  + **Combining data** with the FitBit dataset
  + Video lengths: 15 min + 8 min = 23 min
  + Add 10 minutes = 33 minutes
* Ask if anyone has questions on lab
  + This is kind of how you get started on a project
    - Data exploration first to understand what is in your dataset
* Recap and end for day

# Day 3

M3L2 – Process Train (0900-1100)

* Play beginning video:
  + EDA: <https://youtu.be/QiqZliDXCCg?si=EJHG-NygHm7d4uH8>
* M3L2 – Process 2: Train + Test
  + Titanic Train dataset for outliers and missing values
  + (10+14 +16+17+13 min = 1:15) – give 1.5 hours total
* Talk about what they saw.
  + It’s important to take note of everything you do for training set and do exactly that for test set. Otherwise next steps won’t work.

M3L3 – Analyze-1 (1100 – 1130)

* M3L3 – Analyze Theory 1: (10 min + 15 min)
* Random Forest: <https://www.youtube.com/watch?v=cIbj0WuK41w&ab_channel=Econoscent>
* Next, run lectures (first 2)

M3L3 – Analyze-2 (1130-1215)

* M3L3 – Analyze Theory 2 Lecture (Random Forest and Boosted Trees)

# Day 4

* M3L3 – Analyze Classification 1 (Lab) (33 min = give 45 min to complete)
  + (0900-0945)
* M3L3 – Analysis Classification 2 (Lab) (17 min = give 30 min to complete)
  + (0945-1015)
* M3L3 – Analysis Test (Lab) (24 min = give 45 min to complete)
  + (1015-1100)
* Show Kaggle 100% completion notebook at bottom of page
* M3L3 – Estimation (Lab) (32 min = give 45 min to complete)
  + (1100-1145)
* Quiz – 15 minutes – go over with class (1145-1200)
* Share and Act – 30 min (1145-1215)
* Show that David McCandless Video is great (but don’t go over it)
* **Conclusion**: show DS workflow, talk about all the things we covered in just 4 half days
  + **Use this** to make successful DS projects. The most important parts are the beginning (goal) and end (presentation).