C772 – Course 11 "Data Analytics Graduate Capstone" (3cr)

- **54.3hr/cr – 163.00hr (5.5wk) – 125pg paper/code (156.75hr) – 0.50hr Video Lectures – 5.75hr Textbook & Labs**
- ***Definition:*** *Demonstrate abilities developed as grad student. Integrate skills and knowledge from several program domains into one project.*
- **Task "0" – Course Setup / Plan of Action / Installation:** 6.25 hr (3 d) – 10/18:10/21 – 5.75 hr TextBooks – 0.50 hr Videos
  - DONE – 0.75 hr – Review KickOff Email from course instructor, pulling out links and info (ugg, work laptop disallow pulling down docx files, lost 20min)
  - DONE – 1.25 hr – LINK – Read Course Chatter
  - DONE – 0.50 hr – LINK – Read Latoya's Unofficial Student Advice Guide
  - DONE – 0.75 hr – LINK – Read C772 Tip Sheet (lots of good instructions and links pulled here and listed below)
  - DONE – 0.25 hr – LINK – Watch MSDA Capstone Best Practices Video
  - DONE – 0.25 hr – LINK – Watch Sample Capstone #1 – Lynda Foster, PPT with video
  - DONE – 0.10 hr – LINK – Read Reddit C772 Capstone Notes – Lynda Foster too
  - DONE – 0.50 hr – LINK, QUIZ – Read "Human Subjects FAQ" page and complete Quiz (must use email ID with @student.wgu.edu instead of wgu.edu), print email to PDF
  - DONE – 0.25 hr – LINK, VERIFY, HOSTEDPANOPTO – Read up on Panopto and request license (long lead, do it early) from IT so have it ready for later
  - DONE – 1.75 hr – LINK – Read Course Material, 4 Modules
- **Task 1a – Conduct Research, Develop Prospectus / Research Proposal:** 65.75hr (20d) – 10/21 : 11/09

  Proposal…so use Future Tense.
  - DONE – 0.50 hr – LINK – Read thru Rubric for Task 1, download docs, will return later and break out tasks
  - DONE – 0.50 hr – LINK1, LINK2 – Wikihow – How to find a topic for your research paper, and How to find statistics for paper
  - DONE – 0.25 hr – LINK – John Earnshaw – How to find the best research paper topics + misc links from there (interesting)
  - DONE – 0.50 hr – LINK – University of Michigan – How to select research topic
  - DONE – 0.10 hr – LINK – Kansas State University – How to Develop a Good Research Topic
  - DONE – 0.15 hr – LINK – David Taylor – Choosing and narrowing research topics for API & MLS Essays
  - DONE – 0.00 hr – LINK – Model Template (or google drive LINK) – Already downloaded from Course Chatter (or from kick-off email)
  - DONE – 0.25 hr – LINK – WGU Library – Full text articles (looked thru, maybe use later)
  - DONE – 0.75 hr – LINK – WGU Capstone Archive – Lots of de-identified prior capstones as examples…pull ideas from these, esp. the Excellence Awards
  - DONE – 0.25 hr – LINK – Google "ProQuest" or "Recommendations for Further Research" + healthcare or whatever; "Google Scholar" … Excellent – use below
  - DONE – 13.50 hr – 50+ Links – Research Open Data Sets for Ideas
  - DONE – 0.25 hr – LINK – Fill out "Data Analytics Capstone Release Form"
  - DONE – 0.50 hr – EMAIL – Review all links from mentor's email (previously seen most, but a few new ones to review)
  - DONE – 7.50 hr – LINK – **DRAFT #1** of "Data Analytics Topic Approval Form" - Research and write-up
  - DONE – 0.10 hr – LINK – Review Rubric for Task 1 in great detail, pulling out items and adding below
  - DONE – 0.15 hr – "Instructional Review Board Quiz and Approval Form" (took test earlier, print emailed pass results to PDF and attach)
  - DONE – 0.25 hr – Email Draft to Capstone Instructor prior to submission…include (1) Topic Approval Form, (2) Release Form, and (3) Human Subjects FAQ Quiz results to CI for sign-off
  - DONE – 0.25 hr – LINK – Excellent reference card for Data Viz in R using ggplot2 package
  - DONE – 0.50 hr – LINK – Extract dataset #1 CDC'S Covid Tracker / Underlying Medical Conditions data set (3,142 data rows b/c that many counties in US) … secondary set
  - DONE – 0.25 hr – LINK – Extract dataset #2 CDC'S Conditions Contributing to COVID-19 Deaths data set (12,311 rows)
  - DONE – 0.10 hr – LINK – Extract dataset #3 CDC'S Death Certificate data set (12,311 rows like above, but state level with age buckets…so after I clean, will be fewer rows)
  - DONE – 0.50 hr – LINK1, LINK2 – Extract dataset #4 CDC'S COVID-19 Case Surveillance Public Use data set
  - DONE – 0.15 hr – LINK1, LINK2 – Extract dataset #5 CDC'S Stacks, Estimated County-Level Prevalence Underlying Medical Conditions for COVID-19

- DONE – 0.25 hr – LINK – Extract dataset #6 CDC Lab confirmed Hospitalizations, with all risk factors present including underlying medical conditions
- DONE – 0.25 hr – LINK – Extract dataset #7 Covid Tracking Project
- DONE – 0.25 hr – LINK1, LINK2 – Extract dataset #8 Kaggle NYT dataset
- DONE – 2.50 hr – Explore Bing and Google for any public dataset having underlying medical conditions with Covid
- DONE – 6.75hr – **DRAFT #2** of "Data Analytics Topic Approval Form" - Research and write-up
- DONE – 0.50 hr – Meeting with Course Instructor to review Approval form
- DONE – 24.75 hr – Proof of Concept of Task 2 for Viability. Wrote 850 lines of R Code. Built Random Forest, Decision Tree, Logistic Regression, and MCA models. 90% done with modeling and coding.
- DONE – 2.75 hr – **DRAFT #3** of "Data Analytics Topic Approval Form" - Research and write-up – Email to Course Instructor for Approval
- DONE – 0.50 hr – Submit Both Forms as Task 1 Submission (make sure submit approval form with CI's signature)

- **Task 2a – Finalize Research & Write Report:** 51.25hr (9d) – 11/09 : 11/17
  - *20-60 pages. Use any tools. Includes graphs, charts, and dashboards.*
  - *Follow C744 Project WorkFlow on this one!*
  - DONE – 0.25 hr – LINK – Read thru Rubric for Task 2, break out tasks below, figure out plan of attack (while waiting for initial emailed review and then final review of Task 1)
  - DONE – 3.00 hr – Write **Topic A - Research Question**: Summarize the original real-data research question from Task 1 (justification + description + hypothesis)
  - DONE – 4.50 hr – Write **Topic B - Data Collection**: Report on process, describe relevant data, advantages & disadvantages of methodology used, detail any challenges
  - DONE – 17.25 hr – Write **Topic C, Data Extraction & Preparation**: Describe. Provide screenshots for each step as evidence. Explain tools & techniques – how & why, advantages & disadvantages
  - DONE – 16.00 hr – Write **Topic D, Analysis**: Process, techniques, justification of tools used, calculations performed, and results. Why + advantages/disadvantages.
  - DONE – 7.50 hr – Write **Topic E, Data Summary & Implications**: Summarize implications of analysis. Results. Limitations. Recommend course of action. Propose two directions for future study.
  - DONE – 0.00 hr – Write **Topic F**, In Text Citations and References (collect as go along)
  - DONE – 2.75 hr – Edit Draft, then cross compare against rubric and all example papers
  - DONE – 0.00 hr – Submit Paper as PDF, and zipped up R code and submitted, even though not required
- **Task 3a – Create Multimedia Presentation:** 39.75hr (7d) – 11/18 : 11/24
  - *10-15 minute presentation, 2 or 3 takeaways. Dazzle them with graphics.*
  - DONE – 0.10 hr – LINK – TED Talk – Talk nerdy to me
  - DONE – 0.90 hr – LINK – PowerPoint Tips – Assertion Evidence Approach
  - DONE – 0.10 hr – LINK – Other Sample Graduate Capstone Presentations (saw this previously)
  - DONE – 0.15 hr – LINK – Read thru Rubric for Task 3, download docs, will return later and break out tasks
  - DONE – 12.25 hr – **Write Executive Summary & Implications**: Problem statement, hypothesis, summary data analysis process, outline of findings, limitations of tools & techniques, proposed actions, expected benefits
  - DONE – 20.50 hr – **Write Presentation of Findings**: direct to audience of non-practitioners, capture me talking + **PowerPoint**, intro of me, problem statement/hypothesis, etc. – follow rubric (or xref wgu course material)
  - DONE – 5.00 hr – **Film Presentation of Findings**
  - DONE – 0.75 hr – LINK – Panopto FAQ, upload video, get email back with link to uploaded video, Submit Task 3

Please reach out to me on LinkedIn at www.linkedin.com/in/mpierceAD576 if these Study Guides were helpful. Always happy to hear from folks.

Open Data sets

Best:

- Kaggle (recom'd for Capstone): https://www.kaggle.com/
- U.S. Gov Open Data: https://www.data.gov/ or http://data.gov/

General:

- Amazon Public Datasets: https://aws.amazon.com/opendata
- Census Academy (new): https://www.census.gov/academy?utm_campaign=20190514msccas1ccstanl&utm_medium=email&utm_source=govdelivery
- Census Bureau: http://www.census.gov
- Computational Search Engine: http://wolframalpha.com
- DataCenterHub: https://datacenterhub.org/
- Econ Lib: http://www.econlib.org/library/sourcesUS.html
- Forbes 1: https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#348fcdd3b54d
- Forbes 2: https://www.forbes.com/sites/bernardmarr/2018/02/26/big-data-and-ai-30-amazing-and-free-public-data-sources-for-2018/#6a0609f5f8ae
- GitHub Awesome Public Datasets: https://github.com/awesomedata/awesome-public-datasets
- Google Search Trends: https://trends.google.com/trends/?geo=US
- ~~Google Correlation **Tool**: https://www.google.com/trends/correlate~~
- Purdue: https://purr.purdue.edu/publications
- Re3data.org: https://www.re3data.org/
- Springboard: https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- ~~Top 50 for Data Science Projects: https://blog.journeyofanalytics.com/50-free-datasets-for-data-science-projects/~~
- UC Irvine Machine Learning Repo: https://archive.ics.uci.edu/ml/index.php
- USA Gov: http://www.usa.gov/Topics/Reference-Shelf/Data.shtml
- Yelp Dataset: https://www.yelp.com/dataset

Geographic:

- AggData: https://www.aggdata.com/
- TIGER (roads, RR, rivers, etc.): www.census.gov/geo/www/tiger
- OpenStreetMap: www.openstreetmap.org/

Health:

- Berkley Public Health: http://guides.lib.berkeley.edu/publichealth/healthstatistics/rawdata
- CDC 1: https://www.cdc.gov/nchs/data_access/ftp_data.htm
- CDC 2: https://www.cdc.gov/nchs/ahcd/datasets_documentation_related.htm
- Data Science Central.com: https://www.datasciencecentral.com/profiles/blogs/10-great-healthcare-data-sets
- ~~Global Health Facts: www.globalhealthfacts.org/~~
- HealthData.gov: https://healthdata.gov/search?sort_by=changed
- National Institute of Health: https://www.ncbi.nlm.nih.gov/guide/data-software/

- Public Health Intelligence: http://publichealthintelligence.org/content/resources/data-sources
- World Health Organization: www.who.int/research/en/

Finance:

- OECD Statistics: http://stats.oecd.org/
- World Bank: http://data.worldbank.org/

Scientific:

- ~~ICSU: http://www.icsu.wds.org~~
- Nature: https://www.nature.com/sdata
- Open Science Data Cloud: https://www.opensciencedatacloud.org

Research Topic Sites:

- Google Recommendation for Further: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C48&as_vis=1&q=recommendations+for+further+research+on+covid+predictive+models&btnG=
- IEEE Sample Paper: https://ieeexplore.ieee.org/abstract/document/5686908/
- National Institute of Health: https://www.ncbi.nlm.nih.gov/books/
- National Academies Press: https://www.nap.edu/read/11872/chapter/8

Course Instructor Suggestions for My First Task 1:

Edits and Commentary

1. Project Name. Should be four-to-six words and indicate the type of statistical method used, e.g., “Linear Regression on PPP Loan data.”, “Multivariate Analysis on …….”
2. Topic. Should be four-to-six words. Example: “Predictive Model for PPP Loan Data.”
3. Research Question must be a question that ends with a question mark.
4. Context. Missing statistical study citation.
5. Context. Missing Sentence for “This study will …”
6. Context. Missing Sentence for “The contribution of this study to the field of Data Analytics and the MSDA program is …”
7. Context. Improper citation. Must follow APA-style.
8. Data. Depersonalize. No “I, we, you, us, our …”.
9. Data. Improper citation. Must follow APA-style. Missing period as well.
10. Data. No link to the data is provided.
11. Data. No limitations nor delimitations provided.
12. Data Gathering. Depersonalize.
13. Data Gathering. Missing citation.
14. Data Analytics. Depersonalize.
15. Data Analytics. Design of the study is missing. Must begin with type of normality test.
16. Data Analytics. Missing citation.
17. Justification. Depersonalize.
18. Justification. Improper citation. Must follow APA-style.
19. Project Outcome. . Improper citation. Must follow APA-style.
20. Verify that Instructor signature block is available with all macros.