Wholesale data = demographic data, purchase data

* Demographic data
  + Correlation matrix
  + Heatmap

**CONSOLIDATION/ SUMMARIZING IS CRUCIAL TO RETAINING KEY KNOWLEDGE POINTS**

**A - SOFT SKILLS FOR REMOTE DATA SCIENTISTS**

**B - MINDSET WHEN FACED WITH A PROBLEM TO SOLVE**

1. I am a creative and persistent problem solver
2. I understand the need or problem and why it is important to solve it or what exactly the solution I come up with is expected to do.
   1. I pay and listen attentively to details when others communicate so I can catch inconsistencies to ask purposeful questions to build a firm intelligence of the business: who, what, when, where, why, how and 5 whys.
   2. I have business sense and I am intellectually curious (always asking why and consolidating the response in mind) to probe further on data, responses, assumptions I am skeptical about. I don’t take responses on data at face value only
3. **I am always prepared to give a response for an action or inaction on my part**. How? I give time to deep-thinking a process, problem etc to render smart and useful solutions. I apply the 5 whys till I get to the root cause or simplest version of a concept to be able to explain it clearly in simple language.
   1. I think before I act so that when I act, I have a good well-thought out reason for why and how I did or didn’t do something. It helps me clearly articulate my assumptions, methodology and solution.
   2. I think about the concepts, techniques, tools and my experience which I can leverage to solve the specific problem and provide the desired solution in a smart and efficient way (or leverage skills and expertise from other team members. I know exactly when and how to seek for help).
   3. Data is a **raw material** for discovering as well as **a tool** for solving business problems. Data isn’t a solution to business problems, it’s a tool to allow people to better see, understand, and deliberate on information that the human mind isn’t really equipped to handle.
4. I state the methodology, specific structure of the solution and follow through to completion.
5. I am trained to always communicate using the Pyramid Principle (speak high-level alone but show more instead; always have the high -level response and the detailed response as well when ready) and the Rule of 3 – written, verbal, visual to technical and non-technical audience.
   1. I pitch my data solution and recommendations to stakeholders in a compelling way that inspires stakeholder action and buy-in

Expert Analyst

* Extensively explores data, with speed and accuracy, for interesting insights on:
  + Opportunities (potential products and services to expand customer base, increase revenue and profit)
  + Business problem (pain point that causes business to increase cost, lose profit eg loan defaults erode business capital investment, customer churn where cost of onboarding and projected sales values are affected)
  + This guides formulation of KPQs, which can be tracked by defining metrics and KPIs

Statistician

* Prerequisites for a project:
  + Hypothesis of a business problem to test
  + Stakeholders are interested in using data science to solve the business problem if it is worth solving. The result/ impact of solving the problem is valuable business-wise.
  + Enough and relevant data is available in usable format
* Justifying the need for a project (because you have quantified the claim/ hypotheses raised by expert analyst is worth solving)
  + Don’t rush to dig into data
  + Business Analytics: Understand the business headline which stem from:
    - product (of interest) and it features
    - operations: agent or employee roles, customer acquisition and onboarding, user journey
    - how business is funded
    - business sales
    - industry
  + Understand hypotheses to test
  + Understand and data and build data pipeline from start to finish and iterate multiple times
    - clean data
    - Note assumptions made
    - Clean and transform data (Probability and Statistics)
  + Quantify claimed business problem
    - derive business performance metrics, derive potential candidate features for modeling, derived multiple target outcomes
    - quantify the claimed business problem
    - Currently: loss being made on bad customers or money being left on the table due to not exploring a specific opportunity. Possible savings that could have been made on capital expenditure if that money was not invested in them. That capital could be spent on improving customer experience for existing customers or acquire new good customers
  + Quantify the maximum theoretical impact possible on the portfolio if stakeholders should sign-off on and invest in the project.
* Business Development skills
  + Pitch the need for stakeholders to invest in a project to solve particular question.
    - First state the assumptions made
    - Prove the need for stakeholders to invest in project using statistics and analysis from table

Decision maker

* Decide with stakeholders on details of project:
  + Plan and structure a detailed, written design for program with intended outcome of program
    - **Problem definition:** high-level statement of problem to solve, desired outcome and business value of project eg we want to predict building energy demand accurately 7 days in advance to help building engineers prepare for the next week.
    - **Project requirements:** set of objectives at lower level than the problem statement that solution must meet. Will guide detailed design of program and allow you assess if project is a success eg metric (an error of less than 10%), end-user experience (graphs must clearly show best estimates and uncertainty)
    - **Architectural specification:** Plan files, classes, routines (functions) that will make up my program. Get into details of the design eg error handling, input/output, user interface, data flow through program. At this stage is where you really understand what is being asked of the program and how you will approach the problem. Plan out where you foresee difficulties, make time estimate for the project, outline alternative approaches and assign responsibility during the design phase of project
  + Decide on population of interest
  + Decide on assumptions
  + State logistics etc for project execution, project timelines, stakeholders, deliverables etc. (project team)

Applied Machine Learning Engineer

* Prerequisite
  + Expert analyst, statistician work phase
  + Understand project scope etc as listed from decision maker
* Project kick-off (expose data very well so that algorithm has a good view of data to learn patterns accurately within it)
  + State methodology for solving problem
  + Extract data – internal + external
  + Understand data, preprocess data - clean data, transform data (X, y)
  + Feature engineering
  + Split data and treat any imbalanced classification
  + Baseline model
    - Explain null accuracy terms of customer type count, customer type distribution and initial dumb model performance
  + Run different experiments and take notes.
    - With every experiment, have a clearly defined hypothesis to test
  + Perform economic + business impact analysis of deploying model. Ideal scenario, vs current scenario, model scenario and resulting impact
  + Pitch results to get buy-in. Tie in explanations of with money
  + Deploy selected model in production
  + Maintain model

==================================================================================

WHOLESALE SUPPLY DATA SET PROJECT

SCENARIO:

1. Identify a business problem
   1. Sales prediction
   2. How long do they take between purchases (demand and supply for inventory stocking)
   3. Market basket analysis
   4. Predict purchase amount of customer against various products to help create personalized offers to customers
2. Quantify the business problem with business performance metrics
3. State the maximum theoretical impact of a project in pursuing this project

In a business environment, this means starting with your strategic business objectives and then identifying between 1 and 3 KPQ for each goal. Once you are clear about the questions you need answers to can identify the KPIs. Here again, my advice would be to aim for between one and thee KPIs per question. The overall goal is always to keep numbers of questions and indicators down to the vital few, which generally means the fewer the better. By first designing KPQs we are able to identify our information needs and ask ourselves: ‘What is the best data and management information we need to collect to help us answer our key performance questions?’

PRELIMINARY:

1. How company works

ANALYSIS

1. Acquire domain knowledge on operations of wholesale supply industry
2. Audit data
3. Clean the data after realizing that there is no systemic error in the data worth noting/ correcting
4. Split the data into two – use one to form a theory/ hypothesis and the other to test the hypothesis. If the same phenomenon exists in both datasets, maybe it’s a general phenomenon that also exists wherever those data sets came from.
5. As an expert analyst, I explore the data with speed and accuracy to generate inspiration for KPQs to raise. If the data I have is not sufficient, as a good analyst, I don’t come to conclusions at all (Use statements like I feel, this inspires, I’m not sure). I instead present these interesting but important opportunities and business problem insights worth solving to stakeholders for their decision on which one to pursue further with a statistician. Analytics help me craft good hypotheses (claims) for testing. It improves the quality of the questions I ask, but the patterns I find should not be taken seriously until they are tested statistically on **new data.** Any data point used to inspire a theory or question can’t be used to test that same theory.
6. Decision maker frames decision to be made, sets the decision criteria – sets the decision boundary between the default action and an alternative. They define the metrics, call the shots on statistical assumptions, details of population or users of interest, determine the required level of analytical rigor based on potential impact on the business.
7. Statistics help test hypotheses. It improves the quality of my answers. As a statistician I help to test whether it is sensible to behave as though the phenomenon an analyst found in the current data set applies beyond it. The statistician does the rigorous follow-up of any potential insights the analyst brings up. The statistician is to investigate if the pattern in the data is signal (patterns/facts that exist in your data set and beyond/ real phenomenon that exists in your population of interest) or noise (patterns/facts that exist only in your data set/ idiosyncrasy in my current data set) or apophanies (patterns/facts that exist only in your imagination). Before I start working on the proposed business problem or opportunity I need to ask:
   1. Why the question is being asked (why is this a problem or opportunity)
   2. What theories does the business audience have
   3. What is the business hoping to affect or influence
8. Applied ML Engineer comes up with a system that automates a tricky task well enough to pass a statisticians strict testing bar and delivers the audacious performance a business leader demands. One who is open to leaning from others, responding constructively to feedback on code quality, filling software engineering and data science toolbox with appropriate techniques :
   1. Plan and structure a detailed, written design for program with intended outcome of program
      1. **Problem definition:** high-level statement of problem to solve, desired outcome and business value of project eg we want to predict building energy demand accurately 7 days in advance to help building engineers prepare for the next week.
      2. **Project requirements:** set of objectives at lower level than the problem statement that solution must meet. Will guide detailed design of program and allow you assess if project is a success eg metric (an error of less than 10%), end-user experience (graphs must clearly show best estimates and uncertainty)
      3. **Architectural specification:** Plan files, classes, routines (functions) that will make up my program. Get into details of the design eg error handling, input/output, user interface, data flow through program. At this stage is where you really understand what is being asked of the program and how you will approach the problem. Plan out where you foresee difficulties, make time estimate for the project, outline alternative approaches and assign responsibility during the design phase of project
   2. Build using readable code
   3. Test and review using structural basis testing, pytest, CircleCI
      1. Loading data
      2. Transforming data
      3. Feature engineering
      4. Modeling
      5. Predicting
      6. Storing predictions
      7. Generating model explanations
      8. Validating models
   4. Iterate informed by repeated measurements because fully-formed great solutions are as a result of willingly iterating unto perfection. Eg:
      1. How do you know whether an optimization increased the speed of my code (first code should work before an attempt to optimize)?
      2. How do you know which parts of the code are most error-prone?
      3. How do you know which features spend the most time with?
      4. How do you know which parts of a project take the most time?
   5. Set up a structure to track outcomes over time to evaluate performance in production
   6. Post project – evaluate steps that can be improved to ensure estimate of project timelines are always met

CONCLUSION

1. References

Taking stock of competencies:

DOMAIN KNOWLEDGE

* Expert analytics
  + Understand how business makes money (customers, sales, revenue, suppliers, competitors)
  + Exploring data to come up with KPQs which help define and quantify business problem and maximum theoretical impact of a solution. Have unshakeable curiosity and emotional resilience to handle findings a whole lot of nothing before coming up with something both interesting and important
    - Business toplines (aggregate)
    - Investing based on slicing data on cohort basis eg date of registration etc as it applies to business toplines
  + Defining strategy-based, operations-based KPIs (KPIs that track improvement in processes)
  + Defining KRIs
* Business analytics
  + Business impact analysis
  + Economic value analysis as evidence to inspire recommendations and consequently inspire action for adoption by stakeholders
* Quality data collection and storage pipeline – data devoid of bias
* Feature engineering

SOFTWARE ENGINEERING

* Setting up development environment:
  + OS
  + Project Structure
  + Select code format standard eg black autoformatter
  + Virtual Environment
  + Writing readable codes, oop, and managing python package
  + Code comments show **intent** of code rather than just what the code does writing them while creating the functions, keeping functions as short as possible to perform a single tasks and do it well, function names and parameters define exactly what single task they are meant for (if I write a doc-string using and, I need to re-evaluate if the function is doing a single task and simplify the function), adopt consistent code formatting and naming conventions eg capital letters for constants and state at the top of script, use same error handling method everywhere)
  + Git – version control
  + Webserver to manage applications
  + Automated deployment – API

STATISTICS

* Root cause analysis – 5 whys
* Hypothesis testing and A/B testing
* General statistics and applied statistics

DATA SOURCING

* Web scraping for data augmentation
* Generating fake data using a module
* Simulating Product Usage Data in Pandas (workona, server) + Employee Hiring Data (reading) + Business Simulation in Python (workona, server)
* SQLAlchemy

MACHINE LEARNING ALGORITHMS

* Time Series Analysis
* Customer Analytics
* Cohort Analysis
* Churn Analysis

COMMUNICATION

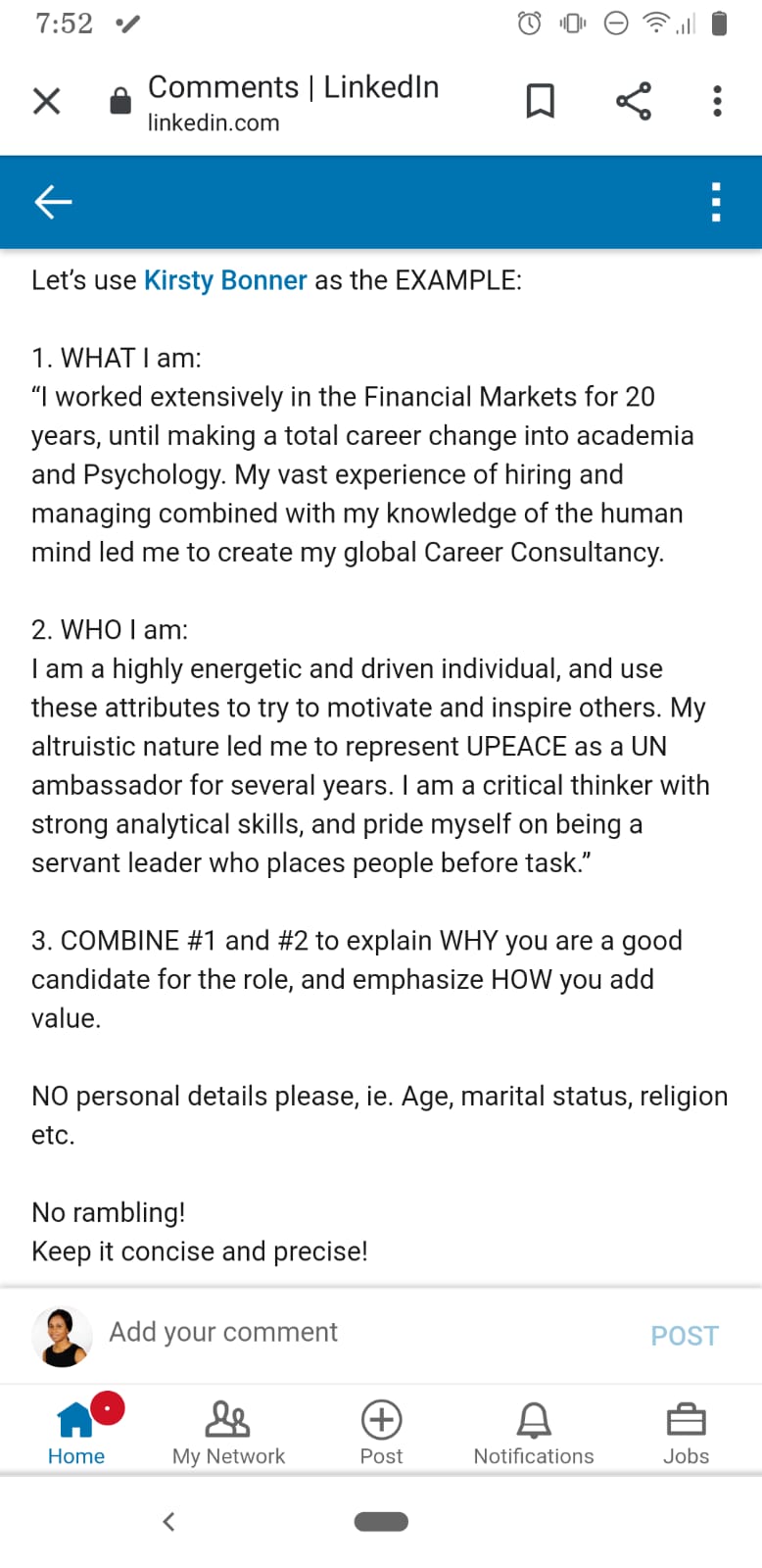
* Data storytelling
  + Crafting life-changing, engaging slides (inspiration from screenshots on phone)
  + Pyramid Principle (also using single syllabus, answering the question directly)
  + How to sell
    - Tell compelling data stories to get stakeholder buy-in and adoption of a data product
* Making recommendations
  + Rule of 3
* Asking purposeful questions

[PROJECT MANAGEMENT](https://docs.google.com/document/d/1rghfOWgiJ_27dDskAWQYwp3SVCTE9sQdW5Zc-tm8sv8/edit?usp=sharing)

* Internal Data Science Sync
  + Weekly
  + Daily
* Client Meetings Slides
  + Pre-kickoff / Pitch
  + Kick-off Meeting
    - Welcome and introduce self
    - Team introduce themselves – name, role, team name
  + Project management logistics and communication tools
    - Github account
    - Slack account
    - Secure file sharing protocol eg FTP, Dropbox, Google Drive
    - Project mailing list
    - Project folders – internal + client (business background, data, deliverables, meeting : recordings + update slides)
    - Project specification document
    - Project progress report
  + Update Meetings (plan for next meeting right after – template of slide, content, dry-runs practice of delivery and answering expected questions at least 2/3 days before actual meeting. Talk less, show more. On meeting day, open necessary links to display)
    - Pleasantries
    - Outline of Presentation
    - Show Timeline and Milestone Tracker
    - Context (brief recap of what was previously done)
    - Whats new: high Level summary of changes made and then show the details. Coordinate what I am saying with what I am saying, speak up and articulately in complete sentences:
      * What has changed per previous meeting
      * Why the change is important
      * What is the business impact (depending on specific client KPI)
    - What’s next
  + Final Presentation Meeting
  + Technical Report

TELL ME ABOUT YOURSELF

* HOW I can add value (managerial or lead role):
  + Decision maker
  + Technical grit with speed and accuracy
  + People skills



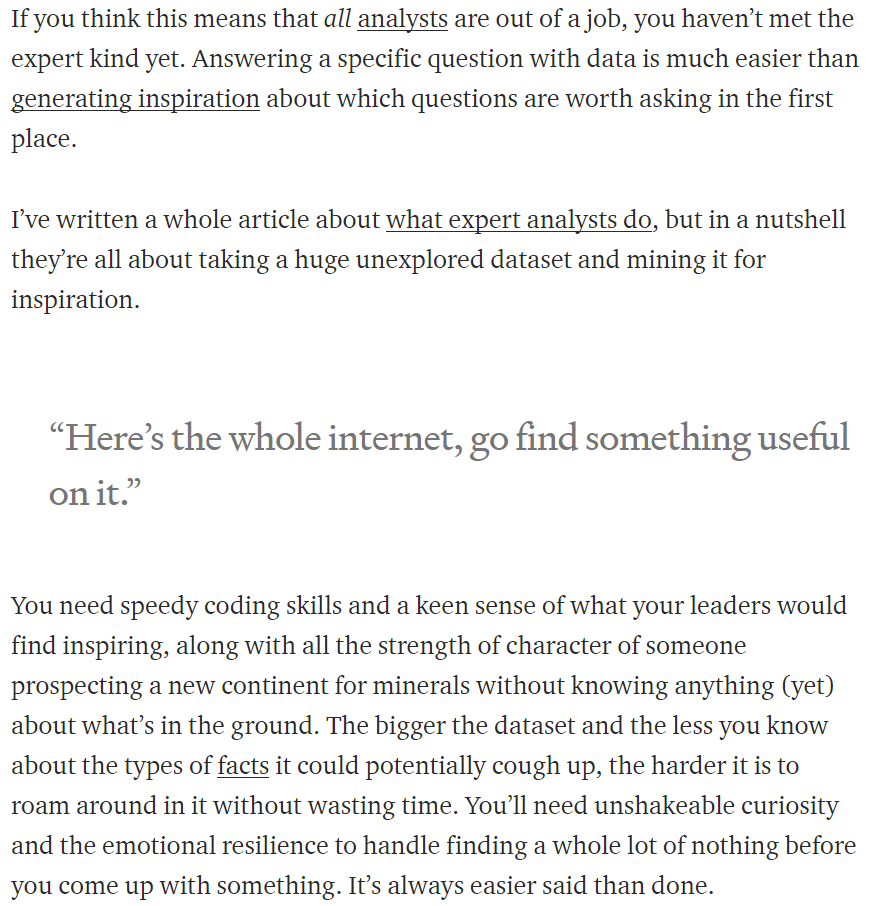
FAVORITE BLOGS

* 365 Data Science
* Analytics Vidhya
* Machine Learning Mastery
* UCananalytics
* Elite Data Science

INFLUENCERS

* Cassie Kozyrkov
* Eric Weber
* Nabih Ibrahim Bawazir
* Srivatsan Srinivasan

Expert Analyst – Inspiration Prospectors



CSF, KAQs, KPQs, KPIs, KRIs – Defining KPIs that inspire action

Interested in KPIs that are more to review and track performance and long-term health of the organization in terms of strategy and not operations: KPQs

365 Data Science

Notes in brown notebook

<https://www.bernardmarr.com/default.asp?contentID=1008>

https://www.bernardmarr.com/default.asp?contentID=1008

https://www.bernardmarr.com/default.asp?contentID=976

https://books.google.com/books?id=tupdDwAAQBAJ&pg=PA58&lpg=PA58&dq=kaqs&source=bl&ots=flqhDESS1i&sig=ACfU3U1HJ6bRPVXZJVEmVSmCwe\_Onrkt9w&hl=en&sa=X&ved=2ahUKEwib5ZzmqITqAhUBca0KHQCuAwQQ6AEwCHoECAUQAQ#v=onepage&q=kaqs&f=false

https://www.datapine.com/kpi-examples-and-templates/

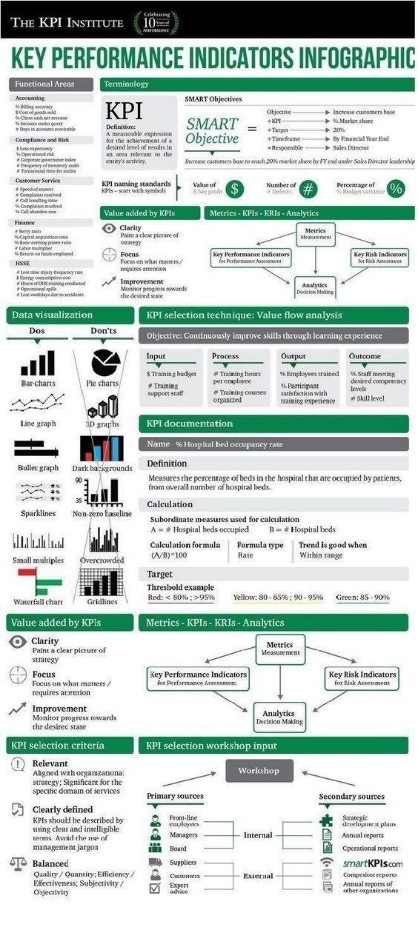
https://cxl.com/blog/key-performance-indicators-kpi/

https://www.linkedin.com/pulse/steve-jobs-approach-developing-kpis-finance-mukogo-ca-insead

https://bernardmarr.com/default.asp?contentID=1406

<https://www.geckoboard.com/blog/defining-kpis-how-to-choose-metrics-that-inspire-action/>

<https://questetra.com/blog/about-kpi-and-kgi/>



From my work experience in the credit risk and lending space, I am comfortable extracting and manipulating big data, building an end-to-end data science pipeline (I am familiar with the sections of the pipeline that should be done in parallel or sequentially to ensure project timelines are met), exploring data to identify and quantify business problems worth solving and opportunities worth exploring, automating tasks using Python, articulating insights and recommendations based off analysis of the business value and economic impact of deploying a model in production. The training in predictive analytics in my Masters program in Applied Statistics was in R, but within 3 to 4 months of joining my current job which happened to be my first full-time data science job, I picked up skills in Python in order to carry out a project requiring the coding to be done in Python. As of today, I have upgraded my skills in Python to the point where I am applying Object Oriented Programming in Python to create a custom package to consolidate scripts my team and I have written and used for multiple projects so we have a central repo of our codes for future projects. I was the first resident data scientist to be employed in my company for the team based of in Ghana, West Africa, and so I have been influential in crafting the playbook and procedures my team carries out in ensuring successful projects. I have also acquired project management skills which makes me well suited for customer facing data science roles. I am a quick learner, intellectually curious, ask purposeful questions to aid my analysis, result-oriented and passionate about delivering value that inspire strategic decisions and actions in every project I am assigned to do

Summary:

1 Data Science Pipeline for Prototyping done in Jupyter Notebook:

- Understand problem

- EDA

- Clean and Transform Data

- Train and Evaluate Model

(All the above in Jupyter Notebook)

2 Set up virtual environment and GitHub account:

- Write production-level code of model (Python script) and push to GitHub from within a virtual environment.

- The model will power the application you will build.

3 Build Flask application with python with gunicorn server (an http server)

- This builds the application

- You will need to write some HTML and JavaScript here.

4 Set up Travis account and link to GitHub account:

- It pulls code from GitHub and automatically runs the tests you've defined on the codes. When codes pass the test it returns zero. This is continuous integration.

5 Set up Herouku account and link to ...:

- When codes pass test in Travis, codes are automatically pushed to a Herouku (a live server). This is continuous delivery

|  |  |  |
| --- | --- | --- |
| Core Skills | Core Tools | Soft Skills |
| * Demonstrate curiosity and an analytical mindset * Forming and testing a hypothesis * Use subject matter knowledge to ask the right questions and identify areas of business impact * Design an experimental project and apply the workflow of data science * Curate data relevant to business outcomes * Project manage a data science project to successful completion * Present a data science project to stakeholders * Explore and analyse data for insights using visualization * Explain insights with clear and compelling written, verbal and visual communication * Curate visualizations that assist stakeholder understanding * Present storybooks using dashboards * Illustrate entity relationships graphically * Apply the fundamentals of languages for effective analysis * Write code to read data, access packages, apply logic * Debugging, profiling and optimization * Organizing and commenting code * Map data sources to data visualization libraries * Write reusable libraries * Clean data through statistical approaches, such as identifying outliers * Transform data into machine-readable formats * Create, read, update and delete on databases and apply data normalization * Extract, Transform and Load data * Build, monitor and measure quality of models over time * Deep knowledge of statistical and mathematical concepts * Identify trends and behaviours with descriptive statistics * Apply inferential statistics on randomised samples to understand characteristics of a population * Apply common models such as linear and logistic regression for prediction * Run A/B tests * Select appropriate algorithms for prediction | Presentation e.g. Excel, Tableau,powerpoint, Project Managements e.g. Trello  Analytics  Notebooks e.g. Jupyter,Google codelab  Data Visualization e.g., Matplotlib, ggplot, Seaborn  Scripting and environments e.g. Python, R,  Matlab, Julia  IDEs, Interpreters and Notebooks  Source control e.g. git, Github  Databases and data stores e.g. MySQL,  Postgres | * Agile Problem-Solving * Business acumen * Creative * Critical thinking * Data Inquisitive * Data Intuition * Self-starting Motivation * Strong Communication * Team building |