

PROJECTWORK

STEPHEN GODFREY

PROJECT PORTFOLIO

Selected projects demonstrating Steve's analytical approach and skill set in capturing, analyzing and extracting insight from data





 From careful listening and insightful questioning, establish a well-understood problem statement, delineate constraints and establish a clear project objective

Tool Development

 Using a wide range of technologies including Excel, Google Sheets and the Python-based data science suite, build a user-friendly, easilymaintained and robust tool set

Data Visualization and Analysis

 Employing an extensive background in quantitative analysis, data science and business management, extract insights from the data

×=

Documentation

 With a straightforward style and an attention to detail, write thorough and understandable documentation with helpful graphics

PROJECT I:

Automated Reconciliation & Accounting Files

USERS

Finance department at a business-to-business payments FinTech startup

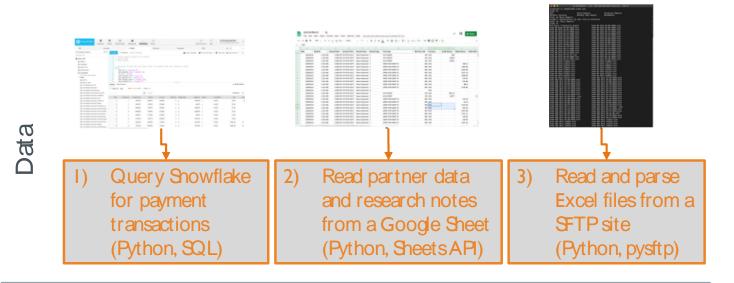
PURPOSE:

Reconcile platform payments with provider reports and generate accounting files

TOOLS& TECHNOLOGY:

Python, Snowflake Connector, SQL, Google Sheets API, pysftp, pandas

WORKFLOW





Results

Reconcile transaction datasets, produce an exceptions report and generate accounting entries (Python, pandas)

5) Output results to a Google Sheet for users to review and import into QuickBooks (Python, Sheets)





PROJECT 2:

TensorFlow Pipeline

USERS

Data scientists evaluating various imageclassification models across several datasets

PURPOSE:

Provide tools to efficiently reparameterize deeplearning models

TOOLS& **TECHNOLOGY:**

Python, Tensor Flow Datasets, Tensor Board, pandas, seaborn, Colab

WORKFLOW

ata

 \cap

Analysis

Results

2)

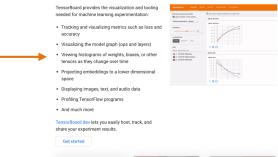
Using a custom wrapper 1) around the Tensor Flow Datasets API. load. display and pre-process ← image data from over 75 classification datasets (TF Datasets)

> Use a generation function to build 3) a model (Python,

Review model results in 5) TensorBoard including example images, model graphs, performance scalars and custom metrics such as a confusion matrices (TensorBoard)

TensorFlow)

	1 TensorFlow	Resources 👻	More 👻	Q Search	English *	GitHub Sign in	
	Datasets						
	Overview Catalog	Guide API					
	TensorFlow	Datasets	a collecti	on of ready	/-to-use data	asets.	
	TensorFlow Datasets is a collection of datasets ready to use, with TensorFlow or other Python ML frameworks, such as Jax. All datasets are exposed as <u>tf.data.Datasets</u> , enabling easy-to-use and high-performance input pipelines. To get started see the guide and our list of datasets.			<pre>import tensorflow.compat.v2 as tf import tensorflow.datasets as tfds # Construct a ff.data.Dataset ds = tfds.load('mnist', split='train', shuffle_f;</pre>			
<u> </u>				<pre># Build your input pipeline ds = ds.shuffle(1024).batch(32).prefetch(tf.data for exemple in ds.take(1): image, label = exemple('image'), example("labe!</pre>			
				RUN IN A CO NOTEBOOK	energie (mage),		
	1		Ç)	TensorFlow D Datasets and Modets	atas	
	Introducing		TensorFlow Datasets		TensorFlow Datasets		
	TensorFlow D	atasets	on GitHub		(TF Dev Summ	it '19)	
	Read on TensorFlow Bi	og	View on GitHub		Watch the video		
eate IsorBo	ard log	S	4)	evalu mod	npile, fi uate th el sor Flo	е	
		1 TensorF	low Resour	es ▼ I Q Si	arch	English • GitHub	
		TensorBoard	d juide				
		Tensor	Board: Ter	sorFlow's \	visualization 1	oolkit	
		T	and data the class for				





PROJECT 3: Data Science Asset Lineage

USERS

Data Science product managers and machine learning platform engineers

PURPOSE:

Provide usage and downstream lineage tracking of data science products delivered as back-end data

TOOLS& TECHNOLOGY:

Python, GitHub, Teradata, Splunk, Google BigQuery and Dataproc and other internal lineage tools

WORKFLOW

ata

 \bigcap

Analysis

Results

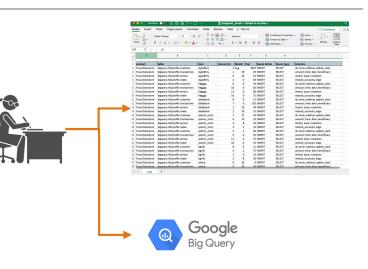
I) Data Science Product Profiles – Output (offline) tables

2) Data Platform Inventory – All platform and database asset inventory

3) Show inventory URLs – Retrieve individual, batch and systemaccount usage of data tables including column detail

4) Git Hub – Find table usage in code 5) Online – Find direct and indirect data usage in online variables, models and edges

6) Report – Save results to BigQuery and create Tableau and Excel user reports



PROJECT 4: Aid-Program Listing

USERS

Small businesses in need of assistance during the COVID-19 pandemic disruption

PURPOSE:

Automate and focus the aid-program search

TOOLS& TECHNOLOGY:

Python, Bing, Scrapy, Beautiful Soup, Selenium, gensim (Doc2Vec), scikitlearn, Google Sheets API (pygsheeets)

WORKFLOW

Data	1)	Bing search scraping - Use a Scrapy spider to search Bing for potentially relevant program URLs			ering I - Elimin unusable progr		
	3)	Sites scraping - Using a Scrapy spider, scrape each program URI for header and body text			ering 2 - Use h iminate progra		
Analysis	5) Scoring - Create document embeddings using Doc2Vec and fit a logistic regression model (with a LASSO penalty) to score each program URL			6) Paring - Eliminate program URLs with scores below a threshold			
Results	7)	W riting - Employ a Google Sheets API wrapper (pygsheets) to write the resulting program listings to 50+ Google Sheets	Name U Program L	SWB_C2CB > States - @ Owner Networks Att Wiscowski AL me Networks Networks me sta Visualization AL me sta Visualization AL me sta Visualization AL me		Image: State	

PROJECT 5:

Compliance Report

USERS

Compliance teams required to regularly report on platform customers

PURPOSE:

Automate the creation of monthly reports

TOOLS& TECHNOLOGY:

Python, Google Sheets API, SQL, pandas, pysftp

WORKFLOW

1)

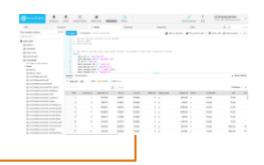
Data

Analysis

Results

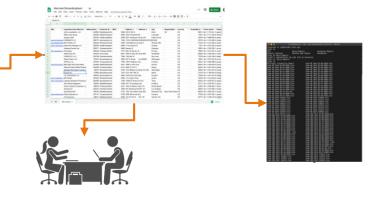
2)

Query Snowflake for customers sending
payments to debit cards and add payment activity and several fields (address, industry, website, etc) for
each customer in a single
SQL statement



Create and store reporting data to establish a report history

3) Output results in Excel for review and posting to the partner's SFTP site (Python, pysftp)



PROJECT 6: Work-program application

WORKFLOW

ata

 \square

Analysis

3)

 Provider status reports – Jobtraining providers upload files with client status reports 2) Case worker input – Case workers input new and update existing client education plans

USERS

Administrators and case workers overseeing a state-funded jobs training program

PURPOSE:

Integrate and report individual and aggregate program metrics

TOOLS& TECHNOLOGY:

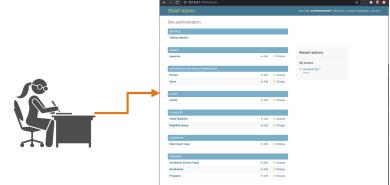
Python, Django, Fake Data, Jupyter



Data-access –
Modifying Django's
model-data-view
base security
model to handle
specific data-access
requirements

- 4) API validation Access Django's REST API framework with Jupyter notebooks to load sample data and validate functionality
- 5) Mock data Use the Faker library to generate mock data to test and demonstrate the application

6) Report – Use the application to manage individual client cases and to administer the program



PROJECT 7: Food Pantry Database Management

USERS

Administrators of an online service to locate food support

PURPOSE:

Curate data to remove inactive pantries and identify low-coverage areas

TOOLS& TECHNOLOGY:

Python, Geopandas, Seaborn, CenPy (US Census data), Haversine

WORKFLOW

Data

Analysis

Results

1)	 Food Support Locations – Geographic coordinates of food pantries US Census – American Community Survey data 			 2) School locations – Geographic coordinates of USpublic and private schools 4) USDA Food Access Research Atlas – USDepartment of Agriculture on food availability 			
3)							
5)	Explore data – Compare pantry data to other sources of food availability and income	Cov Cor data sou	vera mpa a to	pantry ge – e pantry other of food ty	7) Assess pantry prevalence – Measure pantry availability by income and population		
8)	Report – Present findings in presentation desks and Colab Notebooks			Number of Food	d Providers in Database by State		

PROJECT 8:

Automated Inventory Management

USERS

Volunteers providing new shoes to school-age children

PURPOSE:

Reduce manual work to maintain donation inventory

TOOLS& TECHNOLOGY:

Python, Enimoh API into Shopify and Zinc API into Amazon.com, Google Sheets



WORKFLOW

ata

 \square

Analysis

I) Product Parameters – Acceptance criteria (brand, size, price, etc)

2) Product Inventory -Amazon.com shoe availability

3) Goods for sale – Shoes meeting product requirements (URLS) Shopify – Automated loading into a Shopify store

5) Availability – Check of availability on Amazon.com

6) Report – Google Sheets error report identifying Shopify inventory to be fixed



٩	5	e ⊕ 🚏 100% ▼ \$ % .0, .00 123 Defaul ▼ - 10 + B I	÷ <u>A</u>	4
A1		fx		
	٨	8	С	
1		Order Result	Count	
2	0	error	10	
3	1	order_response	1	
4				
5				
6		Order Error Message	Count	
7		The retailer asked for a verification code to verify your account.	5	
8		Zinc or the retailer you requested is experiencing outages. Please try again or contact support@zinc.i		
9		The retailer's final price exceeds the maximum price. Make sure you are passing a value in cents (e.g		
10	3	One of the products you selected is unavailable.	1	
11				
12				
13				
14				
15				
16				
17	_	Offer Explanation	Count	
18		No explanation not available	13	
20		handling davs.max	10	
20		handling_days.max prime != marketplace_fulfiled	10	
22		prime i= markepiace_turned condition in	1	
22	4	contensor_m	1	
24				
25				
25				
27				

PROJECT 9: Polls and Press

WORKFLOW



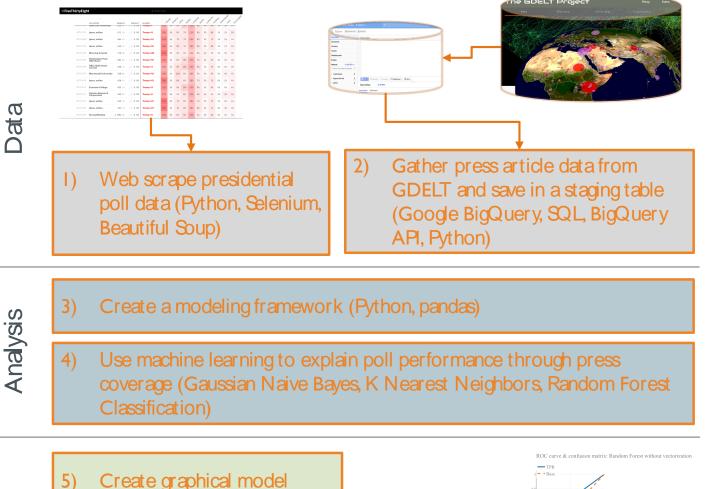
Campaigns and press groups looking to understand drivers of candidate poll results

PURPOSE:

Use press coverage to predict performance in upcoming political polls

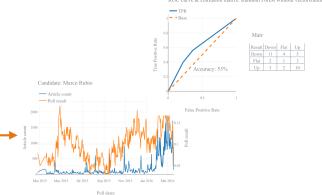
TOOLS& TECHNOLOGY:

Python, Selenium, Beautiful Soup, Google BigQuery & API, SQL, pandas, sci-kit learn



evaluations and output results (Python, Plotly)

Results



PROJECT 10:

An analysis of California's 2018 Camp Fire

USERS

Disaster response teams looking to gain real-time information on the progression of a disaster

PURPOSE:

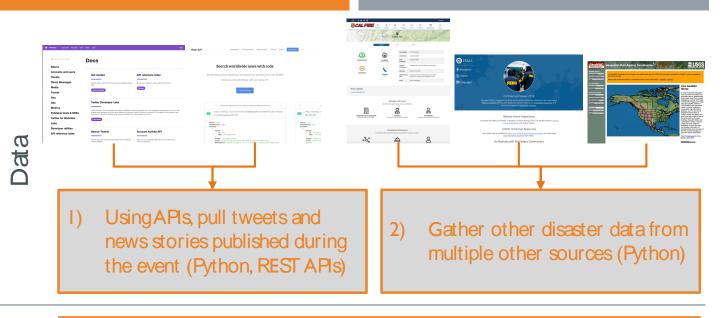
Use real-time data such tweets and news to help direct disaster-response resources

TOOLS& TECHNOLOGY:

Python, Twitter & News APIs, pandas, sci-kit learn WORKFLOW

Analysis

Results



3) Use Natural Language Processing (NLP) techniques to process and categorize text data (Python, pandas, Latent Dirichlet Allocation (LDA), Singular Value Decomposition (SVD), K-Means Clustering)

4) Use regression techniques to correlate Twitter and news data with fire progression metrics (Random Forest Regressors)

5) Generate model results and graphical output (Python, Plotly, ArgGIS, Seaborn, PyShp)

