Credit Card Fraud

This dataset consists of credit card transactions in the western United States. It includes information about each transaction including customer details, the merchant and category of purchase, and whether or not the transaction was a fraud.

Note: You can access the data via the File menu or in the Context Panel at the top right of the screen next to Report, under Files. The data dictionary and filenames can be found at the bottom of this workbook.

Source: Kaggle The data was partially cleaned and adapted by DataCamp.

We've added some guiding questions for analyzing this exciting dataset! Feel free to make this workbook yours by adding and removing cells, or editing any of the existing cells.

Explore this dataset

Here are some ideas to get your started with your analysis...

1. 🗺️ **Explore:** What types of purchases are most likely to be instances of fraud? Consider both product category and the amount of the transaction.
2. 📊 **Visualize:** Use a geospatial plot to visualize the fraud rates across different states.
3. 🔎 **Analyze:** Are older customers significantly more likely to be victims of credit card fraud?

🔍 Scenario: Accurately Predict Instances of Credit Card Fraud

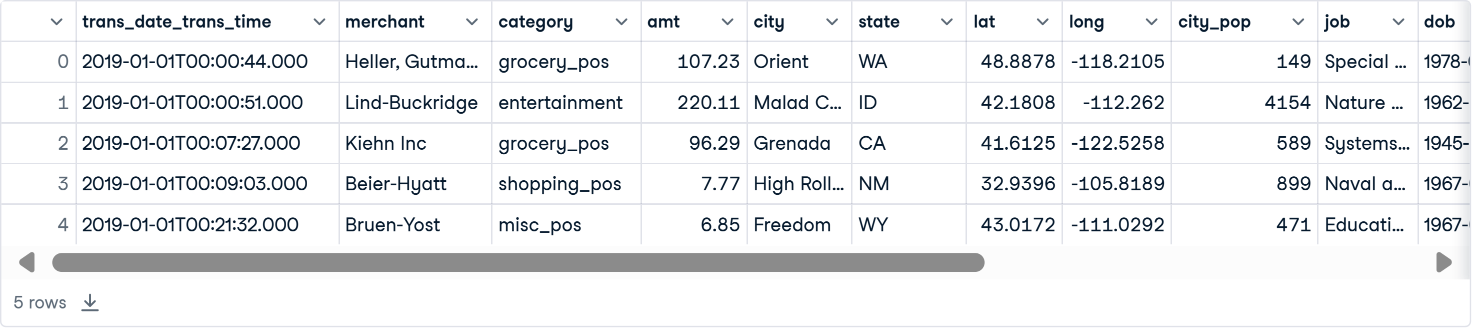
This scenario helps you develop an end-to-end project for your portfolio.

Background: A new credit card company has just entered the market in the western United States. The company is promoting itself as one of the safest credit cards to use. They have hired you as their data scientist in charge of identifying instances of fraud. The executive who hired you has have provided you with data on credit card transactions, including whether or not each transaction was fraudulent.

Objective: The executive wants to know how accurately you can predict fraud using this data. She has stressed that the model should err on the side of caution: it is not a big problem to flag transactions as fraudulent when they aren't just to be safe. In your report, you will need to describe how well your model functions and how it adheres to these criteria.

You will need to prepare a report that is accessible to a broad audience. It will need to outline your motivation, analysis steps, findings, and conclusions.

You can query the pre-loaded CSV file using SQL directly. Here's a **sample query**, followed by some sample Python code and outputs:



	trans_date_trans_time	merchant	category	amt	city	state	lat	long	city_pop	job	dob
0	2019-01-01T00:00:44.000	Heller, Gutma...	grocery_pos	107.23	Orient	WA	48.8878	-118.2105	149	Special ...	1978-
1	2019-01-01T00:00:51.000	Lind-Buckridge	entertainment	220.11	Malad C...	ID	42.1808	-112.262	4154	Nature ...	1962-
2	2019-01-01T00:07:27.000	Kiehn Inc	grocery_pos	96.29	Grenada	CA	41.6125	-122.5258	589	Systems...	1945-
3	2019-01-01T00:09:03.000	Beier-Hyatt	shopping_pos	7.77	High Roll...	NM	32.9396	-105.8189	899	Naval a...	1967-
4	2019-01-01T00:21:32.000	Bruen-Yost	misc_pos	6.85	Freedom	WY	43.0172	-111.0292	471	Educati...	1967-

5 rows

	trans_date_trans_time	merchant	category	amt	city	state	lat	long	city_pop	job	d
13	2019-01-01 00:49:25	Little, Gutman...	shopping_net	83.52	Ravenna	NE	41.0233	-98.9041	2202	Solicitor,...	19
14	2019-01-01 00:56:12	Swaniawski, L...	shopping_pos	317.14	Parks	AZ	35.2563	-111.95	759	Geologi...	19
15	2019-01-01 00:56:59	Reichert, Huel...	shopping_net	113.4	Fort Was...	WY	43.0048	-108.8964	1645	Freight f...	19
16	2019-01-01 01:00:48	Howe Lt	misc_pos	218.71	Littleton	CO	39.5994	-105.0044	320420	Water e...	19
17	2019-01-01 01:02:16	Wolf Inc	grocery_pos	89.11	Meadville	MO	39.7795	-93.3014	964	Tourist i...	19
18	2019-01-01 01:04:48	Vandervort-Fu...	grocery_pos	50.68	Moab	UT	38.5677	-109.5271	9772	Locatio...	19
19	2019-01-01 01:09:41	Ledner-Pfann...	gas_transport	90.54	Hawthor...	CA	33.9143	-118.3493	93193	Editor, ...	19
20	2019-01-01 01:19:02	Schaefer, Mc...	gas_transport	51.33	Manville	WY	42.73	-104.7024	241	Educati...	19
21	2019-01-01 01:22:56	Fisher-Schow...	shopping_net	226.33	June Lake	CA	37.7773	-119.0825	633	Health s...	19
22	2019-01-01 01:23:00	Medhurst PLC	shopping_net	215.99	Sixes	OR	42.825	-124.4409	217	Retail m...	19
23	2019-01-01 01:23:17	Kerluke Inc	misc_net	1.47	Holstein	NE	40.4542	-98.6538	331	Telecom...	19
24	2019-01-01 01:23:50	Bauch-Rayno	grocery_pos	122.05	Westervi...	NE	41.4193	-99.3844	73	Product...	19
25	2019-01-01 01:34:25	Hills-Olson	grocery_net	27.03	Ballwin	MO	38.577	-90.5255	92608	Enginee...	2
26	2019-01-01 01:36:06	Durgan-Aue	misc_net	23.8	Fields La...	CA	40.7268	-124.2174	276	Scientis...	19
27	2019-01-01 01:41:39	Pacocha-Bauch	shopping_pos	2.1	Grenada	CA	41.6125	-122.5258	589	Systems...	19

100 rows

Data Dictionary

transdatetrans_time	Transaction DateTime
merchant	Merchant Name
category	Category of Merchant
amt	Amount of Transaction
city	City of Credit Card Holder
state	State of Credit Card Holder
lat	Latitude Location of Purchase
long	Longitude Location of Purchase
city_pop	Credit Card Holder's City Population
job	Job of Credit Card Holder
dob	Date of Birth of Credit Card Holder
trans_num	Transaction Number
merch_lat	Latitude Location of Merchant
merch_long	Longitude Location of Merchant
is_fraud	Whether Transaction is Fraud (1) or Not (0)

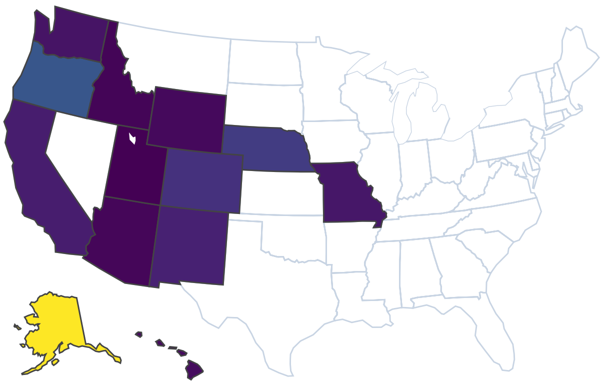
	merchant	is_fraud
0	Stokes, Christiansen and Sipes	
1	Predovic Inc	
2	Wisozk and Sons	
3	Murray-Smitham	
4	Friesen Lt	
5	Raynor, Reinger and Hagenes	
6	Heller-Langosh	
7	Padberg-Welch	
8	McGlynn-Heathcote	
9	Dooley-Thompson	
10	Gottlieb, Considine and Schultz	
11	Moen, Reinger and Murphy	
12	Hauck, Dietrich and Funk	
13	Pouros-Haag	
14	Goyette Inc	

1,782 rows

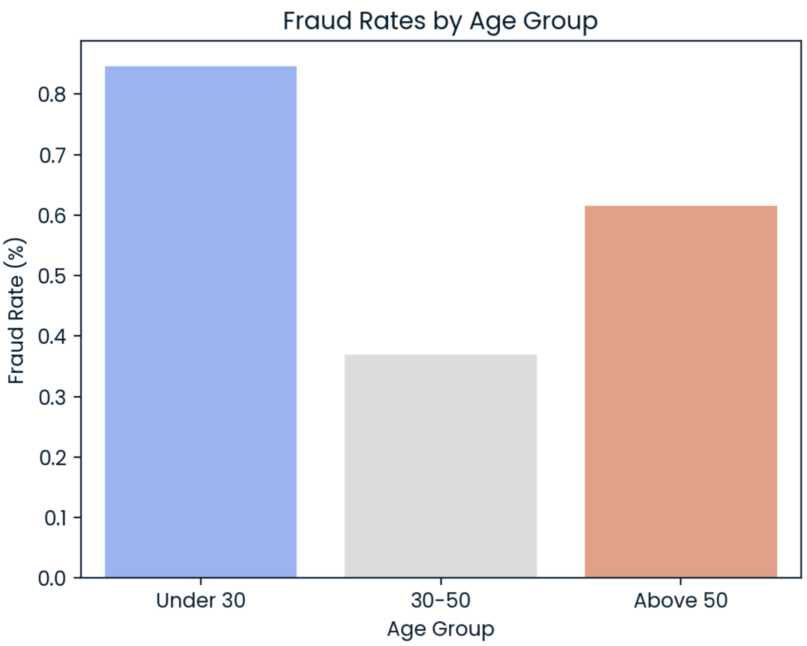
	category	total_transaction
0	shopping_net	
1	misc_net	
2	shopping_pos	
3	grocery_pos	
4	entertainment	
5	misc_pos	
6	home	
7	food_dining	
8	gas_transport	
9	personal_care	
10	kids_pets	
11	health_fitness	
12	grocery_net	
13	travel	

14 rows

Fraud Rates Across States



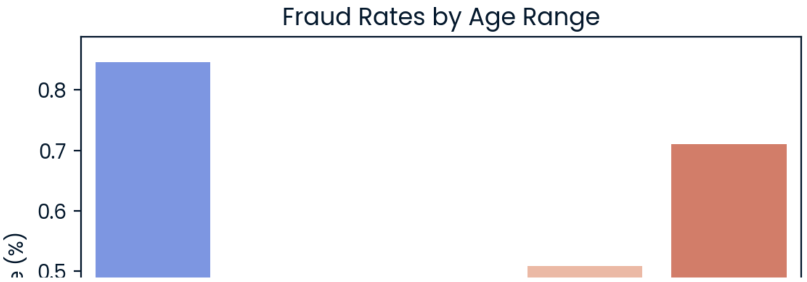
To determine if older customers are significantly more likely to be victims of credit card fraud, you can use statistical analysis combined with visualization.



Statistical Significance Testing To determine if older customers are significantly more likely to be fraud victims, perform a hypothesis test (e.g., Chi-Square Test or T-Test).

To compare fraud rates for specific age ranges, you can calculate fraud rates for custom-defined age ranges and visualize or test them statistically.

Visualize Fraud Rates by Age Range Use a bar plot to compare fraud rates for the specified age ranges.


Fraud Rates by Age Range