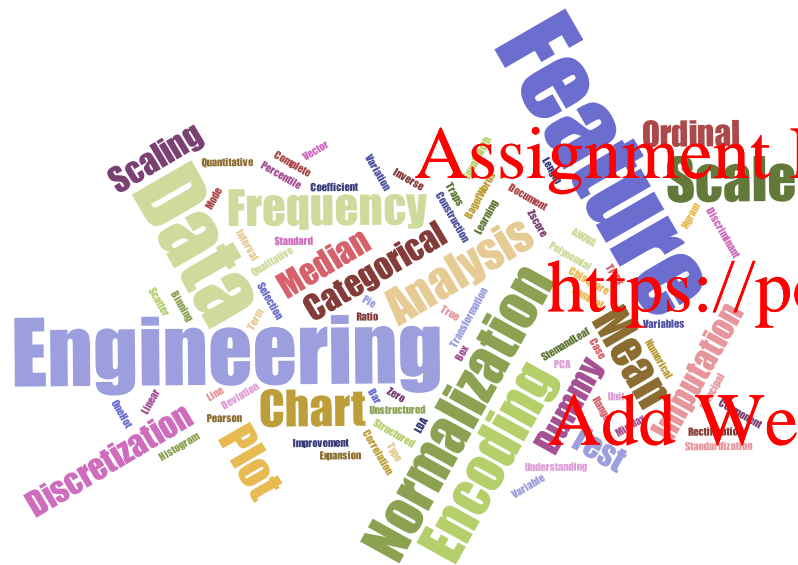


Assignment Project Exam Help

# FEATURE ENGINEERING (CONCEPTS – PART 1)

<https://powcoder.com>  
Add WeChat powcoder



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

# Contents

- Financial Data Sources
- What is Feature Engineering
- Feature Understanding
- Feature Improvement

Assignment Project Exam Help

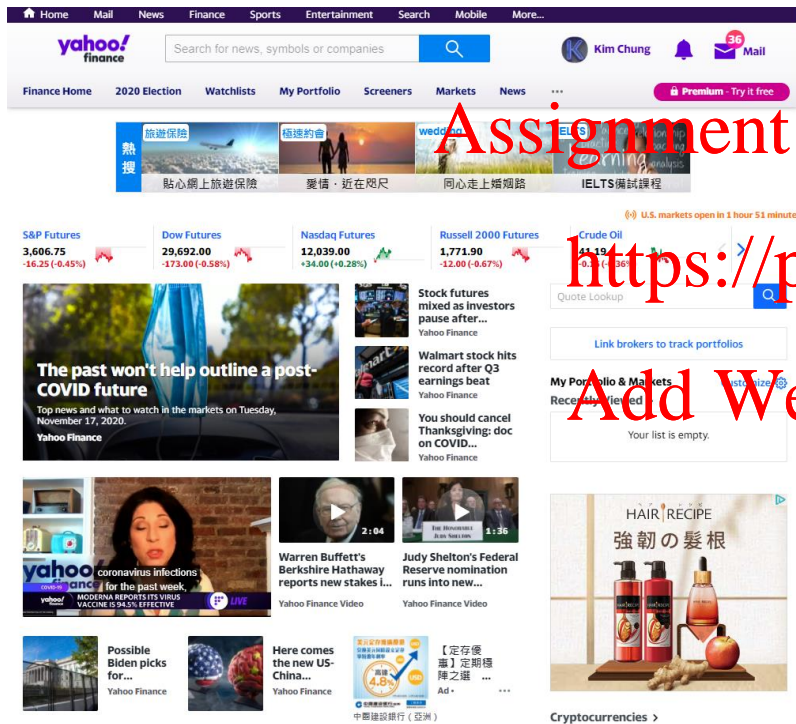
# Financial Data Source

<https://powcoder.com>

Yahoo Finance

Add WeChat powcoder

# Yahoo Finance is one of the reliable sources of stock market data



- Yahoo Finance ([hk.finance.yahoo.com](https://hk.finance.yahoo.com)) supports market summaries, historical & current quotes, news feed about companies
  - Historical & current stock prices in different frequencies (daily, weekly, monthly)
  - Calculated metrics
    - e.g., the beta, a measure of the volatility of an individual asset in comparison to the volatility of the entire market
  - Financial data of a company since its listing in the stock market



Straits Times Index

2,778.55  
+30.55 (+1.11%)

S&P 500

3,616.14  
-10.77 (-0.30%)

Nasdaq

29,797.15  
-153.19 (-0.51%)

Nasdaq

11,912.62  
-1.51 (-0.01%)

Bitcoin USD

11,703.55  
+740.89 (+4.37%)

Singapore markets open in 6 hours 13 minutes

CMC Crypto 200

337.25  
+17.49 (+5.47%)

## Melco Resorts & Entertainment Limited (MLCO)

NasdaqGS - NasdaqGS Real Time Price. Currency in USD

**18.24** -0.28 (-1.54%)

As of 1:47PM EST. Market open.

Add to watchlist

Quote lookup



Summary

Chart

Conversations

Statistics

Historical Data

Profile

Financials

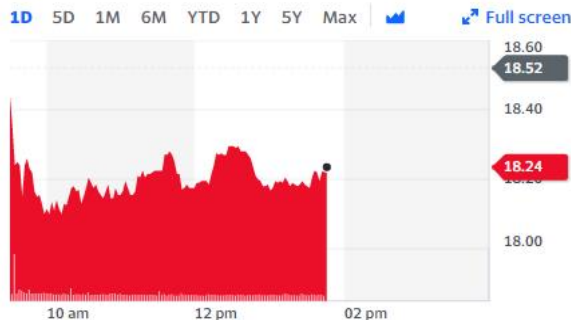
Analysis

Options

Holders

Sustainability

Previous close	18.52	Market cap	1.121B
Open	18.53	Beta (5Y monthly)	1.81
Bid	18.18 x 2200	PE ratio (TTM)	N/A
Ask	18.19 x 1100	EPS (TTM)	-2.09
Day's range	18.09 - 18.53	Earnings date	04 Nov 2020
52-week range	10.81 - 25.22	Forward dividend & yield	N/A (N/A)
Volume	1,249,017	Ex-dividend date	27 Feb 2020
Avg. volume	3,216,324	1y target est	21.45



Trade prices are not sourced from all markets

### 熱門搜尋



一站式室內設計方案

保潔精華



全效保潔精華



速效改善黑眼圈問題

汽車保險



汽車保險即時報價

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

**18.25** -0.27 (-1.43%)

As of 1:56PM EST. Market open.

Summary

Chart

Conversations

Statistics

**Historical data**

Profile

Financials

Analysis

Options

Holders

Sustainability



父母與子女同時投保，子女更可享  
保費半價折扣優惠至18歲

索取報價



**People also watch**

Symbol	Last price	Change	% change
<b>RRR</b> Red Rock Resorts, Inc.	21.25	+0.20	+0.95%
<b>CZR</b> Caesars Entertainment, Inc.	63.41	+0.22	+0.35%
<b>HTHT</b> Huazhu Group Limited	49.94	+2.58	+5.44%
<b>GDEN</b> Golden Entertainment, Inc.	16.36	+0.42	+2.63%
<b>BYD</b> Boyd Gaming Corporation	35.78	+0.14	+0.39%

Time period: 18 Nov 2019 - 18 Nov 2020

Show historical prices

Frequency: Daily

Download

Currency in USD

Date

17 Nov 2020

16 Nov 2020

13 Nov 2020

12 Nov 2020

11 Nov 2020

10 Nov 2020

09 Nov 2020

06 Nov 2020

05 Nov 2020

1D 5D 3M 6M  
YTD 1Y 5Y Max

Start date

01/01/2010

End date

31/12/2020

Done

Cancel

Low	Close*	Adj. close**	Volume
18.09	18.25	18.25	1,271,114
18.33	18.52	18.52	2,636,300
17.30	18.06	18.06	3,734,700
17.06	17.12	17.12	3,670,100
17.31	17.55	17.55	3,316,500
19.13	19.13	18.10	3,553,300
18.20	19.96	18.89	8,124,400
16.30	16.66	16.20	2,411,000
16.88	17.38	16.46	4,157,300

<https://powcoder.com>

Add WeChat powcoder

	A	B	C	D	E	F	G	H
1	Date	Open	High	Low	Close	Adj Close	Volume	
2	4/1/2010	3.47	3.6	3.45	3.58	2.853807	12864900	
3	5/1/2010	3.74	4.3	3.71	4.13	3.292241	23892000	
4	6/1/2010	4.05	4.17	3.81	4.01	3.196583	12123300	
5	7/1/2010	3.95	4.3	3.94	4.3	3.427757	7017700	
6	8/1/2010	4.26	4.35	4.06	4.12	3.28427	6551100	
7	11/1/2010	4.15	4.2	3.92	4.1	3.268327	7246700	
8	12/1/2010	4.01	4.26	3.91	4.01	3.196583	7754300	
9	13/1/2010	4.01	4.01	3.66	3.68	2.933523	16362600	
10	14/1/2010	3.65	3.68	3.48	3.5	2.861779	18437600	
11	15/1/2010	3.64	3.78	3.5	3.75	2.989323	17046600	
12	19/1/2010	3.75	3.76	3.6	3.73	2.97338	3778500	
13	20/1/2010	3.66	3.68	3.56	3.57	2.845835	3787900	
14	21/1/2010	3.71	3.74	3.71	3.63	2.893666	6953300	
15	22/1/2010	3.6	3.61	3.49	3.52	2.805979	3573100	
16	25/1/2010	3.58	3.62	3.37	3.38	2.694377	6726900	
17	26/1/2010	3.35	3.48	3.34	3.39	2.702348	4622600	
18	27/1/2010	3.43	3.49	3.32	3.39	2.702348	2961200	
19	28/1/2010	3.47	3.66	3.44	3.58	2.853807	6747000	
20	29/1/2010	3.68	3.7	3.5	3.57	2.845835	9799000	
21	1/2/2010	3.65	3.85	3.62	3.76	2.997295	11372400	
22	2/2/2010	3.7	4.17	3.31	4.09	3.260355	24943500	
23	3/2/2010	3.96	4.03	3.69	3.79	3.021209	11616600	
24	4/2/2010	3.69	3.72	3.35	3.42	2.726263	6945200	
25	5/2/2010	3.41	3.53	3.3	3.52	2.805979	6984500	

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

*Adjusted closing price adjusted for both dividends and splits*

# Python: Programmatic Access to Financial Data

```
# display the output of plotting commands inline  
# use the "retina" display mode, i.e. to render higher resolution images
```

```
%matplotlib inline  
%config InlineBackend.figure_format = 'retina'
```

Assignment Project Exam Help

<https://powcoder.com>

```
# import the plotting module of the matplotlib package and binds it to the name "plt"  
# display all warnings
```

```
import matplotlib.pyplot as plt  
import warnings
```

Add WeChat powcoder

```
# customize the display style  
# set the dots per inch (dpi) from the default 100 to 300  
# suppress warnings related to future versions
```

```
plt.style.use('seaborn')  
plt.rcParams['figure.dpi'] = 300  
warnings.simplefilter(action='ignore', category=FutureWarning)
```



# Python: Downloading Data as DataFrame

```
# import the relevant packages
```

```
import pandas as pd
import yfinance as yf
```

```
# download the data – data since 1950
```

```
# use "MLCO" as the ticker of "Melco Resorts & Entertainment"
```

```
# disable the showing of the progress bar using "progress=False"
```

```
data = yf.download('MLCO',
                    start='2010-01-01',
                    end='2020-12-31',
                    progress=False)
```

Downloaded 2698 rows of data.

```
# inspect the data using formatted print
```

```
print(f'Downloaded {data.shape[0]} rows of data.')
data.head()
```

	Open	High	Low	Close	Adj Close	Volume
Date						
2009-12-31	3.35	3.48	3.26	3.36	2.678434	7927100
2010-01-04	3.47	3.60	3.45	3.58	2.853807	12864900
2010-01-05	3.74	4.30	3.71	4.13	3.292241	23892000
2010-01-06	4.05	4.17	3.81	4.01	3.196583	12123300
2010-01-07	3.95	4.30	3.94	4.30	3.427757	7017700

Assignment Project Exam Help

# Financial Data Source

<https://powcoder.com>

Quandl

Add WeChat powcoder

# Quandl is a provider of alternative data products for investment professionals

- Quandl delivers **market data** from hundreds of sources via API, or directly into Python, R, Excel, and many other tools
- Featured data includes
  - End of Day US Stock Prices, Core US Fundamentals Data, US Equity Historical & Option Implied Volatilities, Continuous Futures, Trading Economics, BNC Digital Currency Indexed EOD, Global Fundamentals Data, Global Index Prices
- Before downloading data, **create an account** (<https://www.quandl.com>)
- Obtain the **API key** in the profile (<https://www.quandl.com/account/profile>)
- **Search** data function (<https://www.quandl.com/search>)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

# Create a Quandl account

## Start Using Data

Registering for an account provides you with an API key so that you can use our data via all tools, directly through the API and the web interface.

Our platform is used by over 400,000 people, including thousands of analysts from the world's top hedge funds, asset managers and investment banks.

**CREATE FREE ACCOUNT**

**Create your account** STEP 1 OF 3

FIRST NAME  LAST NAME

CHOOSE YOUR PURPOSE FOR USING QUANDL:

☐ Business   
 For a business to access data for a specific, defined use.

☒ Academic   
 Data to be used in an academic environment.

☐ Personal   
 Data for personal use only.

**NEXT →**

**Create your account** STEP 2 OF 3

NAME OF COLLEGE OR UNIVERSITY  SCHOOL EMAIL ADDRESS

HOW WILL YOU BE USING THIS DATA?   
 As an educator

**← PREVIOUS** **NEXT →**

**Quandl** Account Activation

Dear Daniel,  
Welcome to Quandl! Thank you for creating an account.

Please confirm your email address to activate your account.

**CONFIRM ADDRESS**


If you did not create an account, please ignore this email.

Sincerely,  
The Quandl Team  
[connect@quandl.com](mailto:connect@quandl.com)

**Create your account** STEP 3 OF 3

CREATE A PASSWORD  CONFIRM YOUR PASSWORD

☒ I have agreed to the [terms of service](#) and [privacy policy](#)

☒ I'm not a robot  [Privacy](#) [Terms](#)

**← PREVIOUS** **CREATE ACCOUNT**

## Welcome to your new home page.

From here you can view your subscriptions and browse data of interest.

CLOSE X

### NEW PRODUCTS

#### FX Spot Flow Data

FX spot flow data for 33 currency pairs, aggregated in 5-minute intervals and updated daily.

NEW

PREMIUM

HAS SAMPLE DATA



PUBLISHED BY CLS

#### FX Swap and Forward Flow

FX swap and forward flows for 33 currency pairs. Aggregated hourly.

NEW

PREMIUM

HAS SAMPLE DATA



PUBLISHED BY CLS

#### Alternative Data

Explore our curated catalog of institutional-only data products.

REQUEST ACCESS

LEARN MORE

#### Your Organization

You don't currently belong to an organization on Quandl. Learn more about team collaboration and data sharing features.

LEARN MORE

#### Your Data Subscriptions

You are not currently subscribed to any data feeds. Your subscriptions will appear here when available.

# Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## BROWSE

## FILTERS

- ☐ Premium
- ☒ Free

## ASSET CLASS

- ☒ Equities
- ☐ Currencies
- ☐ Interest Rates & Fixed Income
- ☐ Options
- ☐ Indexes
- ☐ Mutual Funds & ETFs
- ☐ Real Estate
- ☐ Venture Capital & Private Equity
- ☐ Economy & Society
- ☐ Energy
- ☐ Agriculture
- ☐ Metals
- ☐ Futures
- ☐ Other

## DATA TYPE

- ☒ Prices & Volumes
- ☐ Estimates
- ☐ Fundamentals
- ☐ Corporate Actions

Equities X Prices & Volumes X Free X Asia X [Clear All](#)

There are 2 databases with data on 'MLCO'

## Hong Kong Exchange

Hong Kong Exchange stock prices, historical data, futures, etc. updated daily.

FREE

HKEX / 06883

[Melco Crown \(06883\)](#)

HKEX / 00200

[Asia Infrastructure Dev \(00200\)](#)

HKEX / 00328

[Alco Holdings \(00328\)](#)

## Bombay Stock Exchange

End of day prices, indices, and additional information for companies trading on the Bombay Stock Exchange in India.

FREE

BSE / BOM524594

[ASHOK ALCO-CHEM LTD. EOD Prices](#)

BSE / BOM507526

[ASSOCIATED ALCOHOLS & BREWERIES LTD. EOD Prices](#)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## &lt; Melco Crown (06883)

## FROM THE DATA PRODUCT:

## Hong Kong Exchange

(43,980 datasets)

## REFRESHED

5 years ago, on 5 Jul 2015

## FREQUENCY

Daily

## DESCRIPTION

Stock Prices for Melco Crown from the Hong Kong Stock Exchange. Currency:

## VALIDATE ⓘ

[http://www.hkex.com.hk/eng/invest/company/quote\\_page\\_e.asp?WidCoID=06883&WidCoAbbName=&Month=1&langcode=e](http://www.hkex.com.hk/eng/invest/company/quote_page_e.asp?WidCoID=06883&WidCoAbbName=&Month=1&langcode=e)

## PERMALINK ⓘ

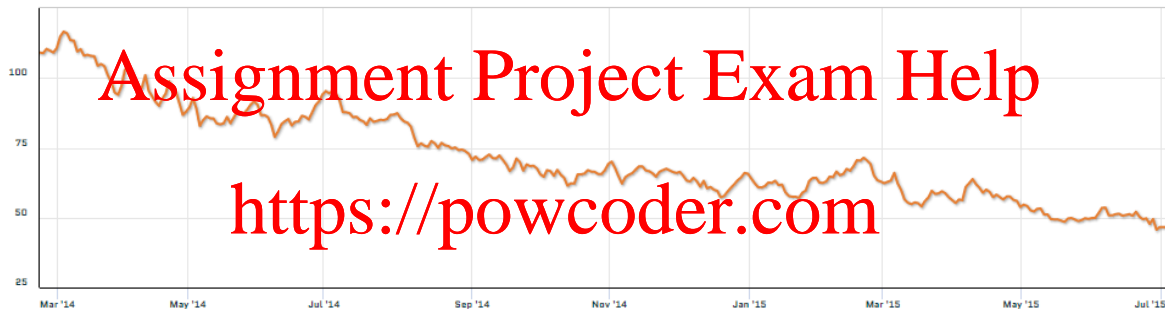
<https://www.quandl.com/data/HKEX/06883>

BOOKMARK

DOWNLOAD ▼

CHART

TABLE

☐ NOMINAL PRICE ☐ NET CHANGE ☐ CHANGE (%) ☐ BID ☐ ASK ☐ P/E(X) ☐ HIGH ☐ LOW ☐ PREVIOUS CLOSE ☐ SHARE VOLUME ('000) ☐ TURNOVER ('000) ☐ LOT SIZEZoom       

From Feb 21, 2014 To Jul 3, 2015

Default ▼

No Transform ▼

## LATEST VALUES

DATE  
2015-07-03NOMINAL PRICE  
46.65NET CHANGE  
0CHANGE (%)  
0BID  
N/AASK  
N/AP/E(X)  
16.20HIGH  
N/ALOW  
N/APREVIOUS CLOSE  
46.65SHARE VOLUME ('000)  
0TURNOVER ('000)  
0

Quandl Code ⓘ

HKEX/06883

## EXPORT DATA

API

## Libraries

## Tools

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Date	Nominal Price	Net Change	Change (%)	Bid	Ask	P/E(x)	High	Low	Previous Close	Share Volume ('000)	Turnover ('000)	Lot Size	
2	3/7/2015	46.65	0	0			16.2			46.65	0	0	300	
3	2/7/2015	46.65	0	0			16.2			46.65	0	0	300	
4	30/6/2015	46.65	0	0			16.2			46.65	0	0	300	
5	29/6/2015	45.7	3.8	7.677	45.65	45.7	15.87	47.5	45.4	49.5	1823	84390	300	
6	26/6/2015	49.5	1.7	3.556	49.35	49.45	17.19	49.8	48.7	47.8	429	21166	300	
7	25/6/2015	47.8	2.1	4.208	47.75	47.8	16.6	49.25	47.75	49.9	561	27069	300	
8	24/6/2015	49.75	0.3	0.100	49.75	49.8	17.21	50.49	47.8	49.45	141	7038	300	
9	23/6/2015	49.45	0.85	1.64	49.4	49.45	17.17	50.75	49.15	50.3	148	7310	300	
10	22/6/2015	50.35	1.75	3.359	50.3	50.35	17.48	50.45	49.85	52.1	46	2303	300	
11	19/6/2015	52.1	1.25	2.458	52.15	52.25	18.09	52.25	51.05	50.85	155	8021	300	
12	18/6/2015	50.7	0.55	1.01	50.7	50.85	17.61	51	50.25	51.35	27	1356	300	
13	17/6/2015	51.25	0.3	0.58	50.95	51.3	17.8	51.3	50.5	50.55	44	2217	300	
14	16/6/2015	50.95	0.55	1.091	50.9	50.95	17.69	51.15	50.85	50.4	32	1617	300	
15	15/6/2015	50.7	0.7	1.362	50.35	50.7	17.61	50.85	50	51.4	24	1223	300	
16	12/6/2015	51.4	0.25	0.489	51.35	51.5	17.81	51.35	51.1	51.15	30	1536	300	
17	11/6/2015	51.25	0.45	0.88	51.5	51.2	17.8	51.5	50.75	50.5	20	1009	300	
18	10/6/2015	50.8	0.05	0.098	50.65	50.8	17.64	50.85	50.6	50.85	24	1202	300	
19	9/6/2015	50.85	2.75	5.131	50.8	50.85	17.66	51.9	50.6	53.6	121	6169	300	
20	8/6/2015	53.6	0.1	0.187	53.55	53.85	18.61	54	53.35	53.5	109	5830	300	
21	5/6/2015	53.5	1.9	3.682	53.35	53.5	18.58	54.6	53	51.6	231	12369	300	
22	4/6/2015	51.6	1.7	3.407	51.55	51.75	17.92	52	50.2	49.9	127	6532	300	
23	3/6/2015	49.9	0.05	0.1	49.85	49.9	17.33	50.25	49.7	49.85	37	1824	300	
24	2/6/2015	49.85	0.3	0.605	49.55	49.85	17.31	49.85	49.15	49.55	33	1647	300	
25	1/6/2015	49.55	0.25	0.502	49	49.55	17.21	50.5	49.65	49.8	5	270	300	



# Python: Programmatic Data Access

```
# import the relevant packages
```

```
import pandas as pd
import quandl as qd
```

```
# use the API key generated during account creation
# provide the DATASET/TICKER of the dataset
```

```
qdl_api_config_api_key = 'GEM.....xWB'
data = qd.get('HKEX/06883',
              start='2010-01-01', end='2020-12-31')
```

```
# inspect the data
```

```
print(f'Downloaded {data.shape[0]} rows of data.')
data.head()
```

Downloaded 332 rows of data.

	Nominal Price	Net Change	Change (%)	Bid	Ask	P/E(x)	High	Low	Previous Close	Share Volume ('000)	Turnover ('000)	Lot Size
Date												
2014-02-21	109.1	0.5	0.460	108.8	109.2	55.15	109.5	108.9	108.6	20.0	2175.0	300.0
2014-02-24	108.9	0.2	0.183	108.9	109.5	55.05	110.0	108.5	109.1	25.0	2751.0	300.0
2014-02-25	110.4	1.5	1.377	110.3	111.2	55.81	112.5	110.2	108.9	43.0	4737.0	300.0
2014-02-26	109.8	0.6	0.543	109.7	110.1	55.51	110.3	109.6	110.4	20.0	2242.0	300.0
2014-02-27	109.1	0.7	0.638	108.7	109.1	55.15	109.5	108.6	109.8	31.0	3333.0	300.0

Assignment Project Exam Help

# What is Feature Engineering

<https://powcoder.com>

Add WeChat powcoder

# Feature Engineering

Assignment Project Exam Help

Feature engineering is the process of transforming data into features that better represent the underlying problem, resulting in improved machine learning performance.

<https://powcoder.com>

Add WeChat powcoder

# Feature engineering is about making data meaningful to the machine learning model

## Raw and Partially Processed Data

Feature engineering can be applied to data at **any stage** and deals with **raw & partially processed data** typically in the form of **classifications (rows) and attributes (columns)**.

## Meaningful Features

A feature is an attribute of data that's **meaningful** to the machine learning process. Some attributes can be unhelpful or even hurtful to the machine learning process.

## Better Representation

Data always serve to represent a specific problem in a specific domain. The rationale is to **transform** data so that it **better represents the bigger problem** at hand.

## Model Performance Improvement

The eventual goal of feature engineering is to obtain data that the learning algorithms will be able to extract patterns from and use in order to obtain **better results**.

Raw data are often in a state that cannot be directly consumed by machine learning algorithms

Assignment Project Exam Help  
<https://powcoder.com>  
Add WeChat powcoder

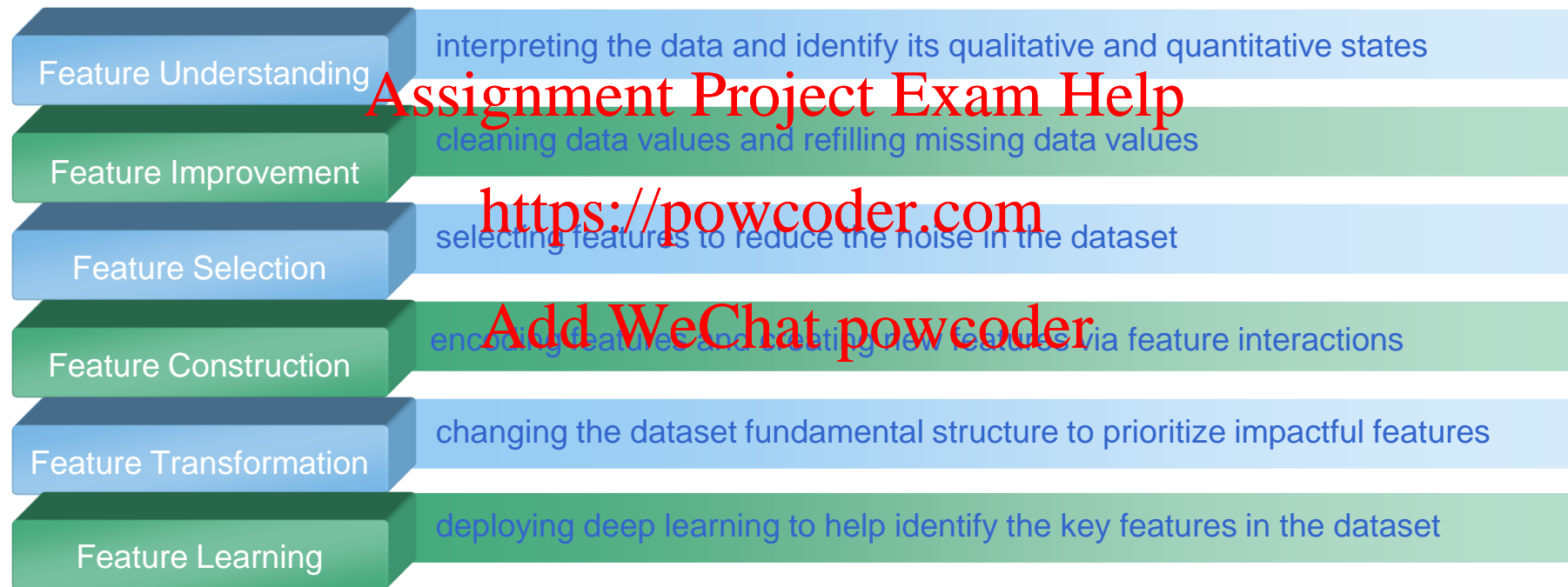
Observations

Name	Amount	Date	Issued In	Used In	Age	Education	Fraud?	Membership	...
Daniel	\$2,600.45	1-Jul-2020	HK	HK	22	Secondary	No	Silver	...
Alex	\$2,294.38	1-Oct-2020	HK	RUS	None	Postgraduate	Yes	Silver	...
Adrian	\$1,003.30	3-Oct-2020	HK		25	Graduate	Yes	Bronze	...
Vicky	\$8,488.32	4-Oct-2020	JAPAN	HK	64	Graduate	No	Gold	...
Adams	¥20000	7-Oct-2020	AUS	JAP	58	Primary	No	Silver	...
...	...	...	...	...	...	...	...	...	...
Jones	₹3,250.11	Nov 1, 2020	HK	RUS	43	Graduate	No	Silver	...
Mary	₹8,156.20	Nov 1, 2020	HK	N/A	27	Graduate	Yes	Gold	...
Max	€7475.11	Nov 8, 2020	UK	GER	32	Primary	No	Premium	...
Peter	₹500.00	Nov 9, 2020	Hong Kong	RUS	0	Postgraduate	No	Bronze	...
Anson	₹7,475.11	Nov 9, 2020	Hong Kong	RUS	20	Postgraduate	Yes	Gold	...

Feature

Target

# Feature engineering can be carried out in steps but different schools have different thoughts on structuring the steps



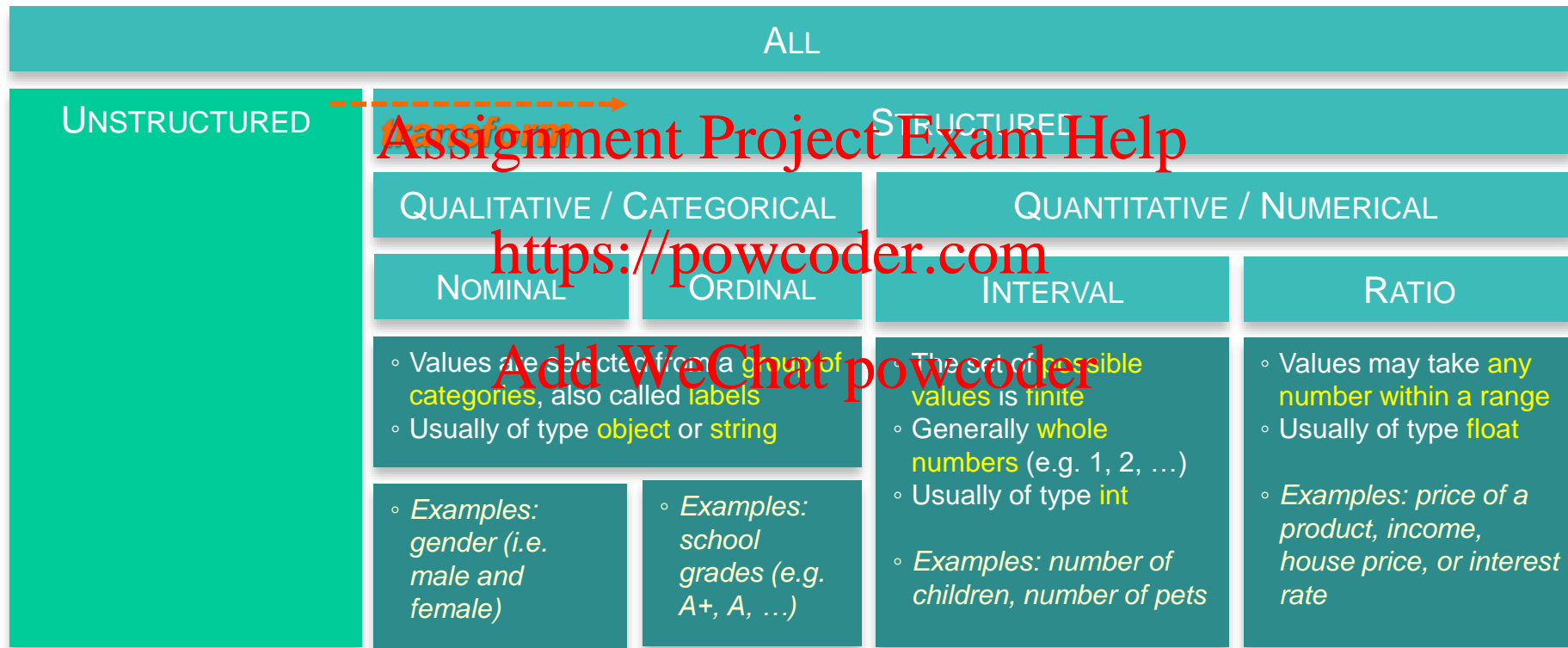
Assignment Project Exam Help

Feature Understanding

<https://powcoder.com>

Add WeChat powcoder

# Correctly identifying numerical & categorical variables involves looking at the data types & inspecting their values





# Understanding the Structure of Data

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## Structured Data

- Any data that can be stored, accessed, and processed in the form of **fixed format**
- The format is known **in advance**
  - e.g. data stored in a **relational database**
- Also referred to as **Schema-on-Write**

## Semi-Structured Data

- Semi-structured data has a **lack of fixed, rigid structure**
- There is no separation between the **data** and the **schema** - a **self-describing structure**
- e.g. XML files, JSON files, web pages in **HTML**, **RDF** files

## Unstructured Data

- Any data with **unknown form**
  - e.g. heterogeneous data sources containing simple **text files**, **images** & **videos**
- Also referred to as **Schema-on-Read**

# Data in a relational database or an Excel spreadsheet are considered structured data

STORE		
Store_key	City	Region
1	New York	East
2	Chicago	Central
3	Atlanta	East
4	Los Angeles	West
5	San Francisco	West
6	Philadelphia	East
.	.	.
.	.	.

PRODUCT		
Product_key	Description	Brand
1	Beautiful Girls	MKF Studios
2	Toy Story	Wolf
3	Sense and Sensibility	Parabuster Inc.
4	Holiday of the Year	Wolf
5	Pulp Fiction	MKF Studios
6	The Juror	MKF Studios
7	From Dusk Till Dawn	Parabuster Inc.
8	Hellraiser: Bloodline	Big Studios
.	.	.
.	.	.

SALES_FACT				
Store_key	Product_key	Sales	Cost	Profit
1	6	2.39	1.15	1.24
1	2	16.7	6.91	9.79
2	7	7.16	2.75	4.40
3	2	4.77	1.84	2.93
5	3	11.93	4.59	7.34
5	1	14.31	5.51	8.80
.	.	.	.	.
.	.	.	.	.

Data in a Relational Database

Month	Name	Gender	Diagnosis	Treatment
May	Jessica	F	Allergy	Eye Drops
May	Sam	M	Allergy	Eye Drops
May	Wes	M	Cataract	Cataract Surgery
May	Rachel	F	Pterygium	Eye Drops
May	Lily	F	Allergy	Eye Drops
May	Hannah	F	Cataract	Cataract Surgery
May	Denise	F	Allergy	Eye Drops
May	Sharon	F	Allergy	Eye Drops
May	Robin	F	Allergy	Eye Drops
May	Lianna	F	Pterygium	Eye Drops
May	Thomas	M	Presbyopia	Reading Glasses
May	Kimberly	F	Refractive Error	Distance Glasses
May	Michael	M	Refractive Error	Distance Glasses
May	Jacob	M	Conjunctivitis	Eye Drops
June	John	M	Presbyopia	Reading Glasses
June	Tim	M	Refractive Error	Distance Glasses
June	Allison	F	Cataract	Cataract Surgery
June	Laura	F	Pterygium	Eye Drops
June	Scott	M	Cataract	Cataract Surgery
June	Sarah	F	Pterygium	Eye Drops
June	Alex	M	Pterygium	Eye Drops
June	Robert	M	Cataract	Cataract Surgery

Data in an Excel Spreadsheet

# Semi-structured data embeds information about the data structure and the data contents in the same document

```
<HTML>
<HEAD>
<TITLE>Your Title Here</TITLE>
</HEAD>
<BODY BGCOLOR="FFFFFF">
<CENTER><IMG SRC="clouds.jpg" ALIGN="BOTTOM"> </CENTER>
<HR>
<a href="http://somegreatsite.com">Link Name</a>
is a link to another nifty site
<H1>This is a Header</H1>
<H2>This is a Medium Header</H2>
Send me mail at <a href="mailto:support@yourcompany.com">
support@yourcompany.com</a>.
<P> This is a new paragraph!
<P> <B>This is a new paragraph!</B>
<BR> <B><I>This is a new sentence without a paragraph break, in bold
italics.</I></B>
<HR>
</BODY>
</HTML>
```

Web Page in HTML format

```
{
  "quiz": {
    "score": 0,
    "q1": {
      "question": "Which one is correct team name in NBA?",
      "options": [
        "New York Bulls",
        "Huston Rocket"
      ],
      "answer": "Huston Rocket"
    }
  }
}
```

Data-Value Pairs in JSON format

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Unstructured data are not really unstructured but structured in a way that is less convenient to manipulate



Assignment Project Exam Help

<https://powcoder.com>

Image in JPEG format



Add WeChat powcoder

Audio in WAVE

# Most unstructured data can be transformed into structured data through a few manipulations

Images are considered **unstructured** data

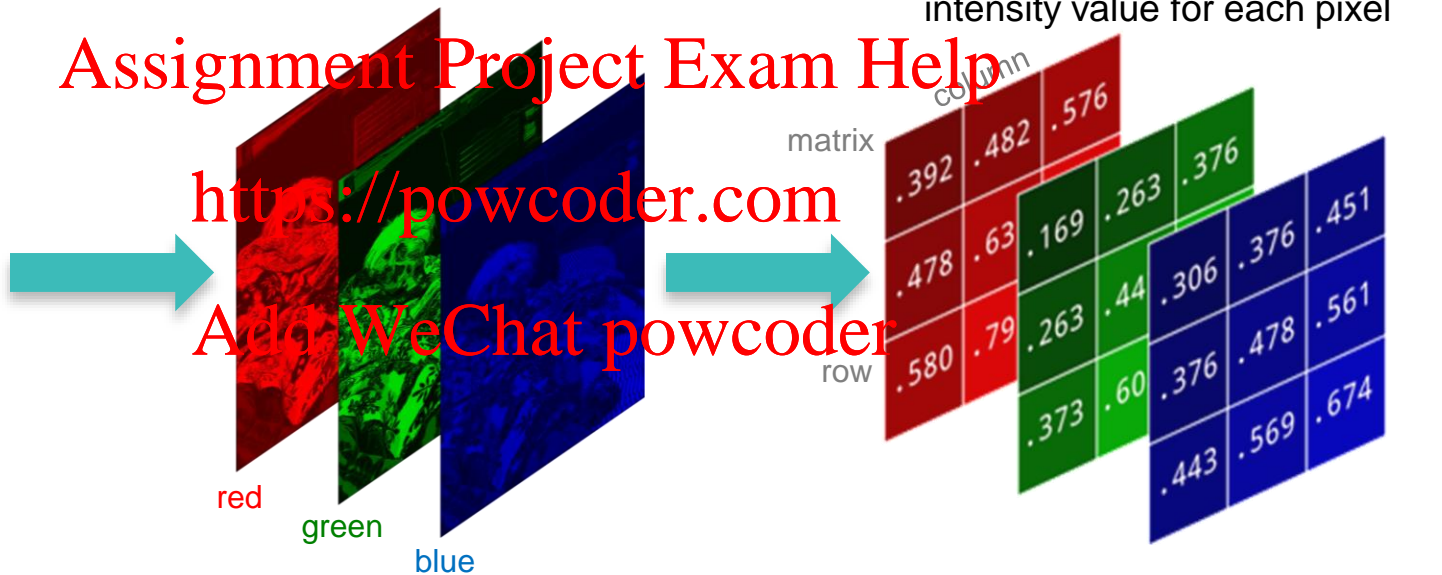


Images can be decomposed into 3 color **channels**

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder







Assignment Project Exam Help

<https://powcoder.com>



Add WeChat powcoder

Training autonomous cars using semantic segmentation of road scenes

# CSV (Comma-Separated Values) Format



Text File

Assignment Project Exam Help



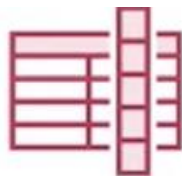
Used to represent tabular structure

<https://powcoder.com>



Add WeChat powcoder

Each line is a record



Each record has multiple columns separated by comma



Assignment Project Exam Help

Feature Improvement

<https://powcoder.com>

Add WeChat powcoder

# Feature improvement is about altering data values and removing dataset columns/rows

- Feature improvement involves both feature **cleaning** and **removal**
  - Cleaning alters columns and rows in the dataset
  - Removal takes columns and rows away from the dataset
- Possible actions include
  - Identifying missing values
  - Removing harmful data
  - Imputing (filling in) missing values
  - Normalizing/standardizing data
    - Z-score normalization, min-max scaling, L1 & L2 normalization

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

# Feature understanding focuses on data values and data value induced structural changes

Observations

Name	Amount	Date	Issued In	Used In	Age	Education	Fraud?
Daniel	\$2,600.45	1-Jul-2020	HK	HK	22	Secondary	No
Alex	\$2,294.38	1-Oct-2020	HK	RUS	None	Postgraduate	Yes
Adrian	\$1,003.30	3-Oct-2020	HK		25	Graduate	Yes
Vicky	\$8,488.32	4-Oct-2020	JAPAN	HK	64	Graduate	No
Adams	¥20000	7-Oct-2020	AUS	JAP	58	Primary	No
...	...	...	...	...	...	...	...
Jones	₹3,250.11	Nov 1, 2020	HK	RUS	43	Graduate	No
Mary	₹8,156.20	Nov 1, 2020	HK	N/A	27	Graduate	Yes
Max	€7475.11	Nov 8, 2020	UK	GER	32	Primary	No
Peter	₹500.00	Nov 9, 2020	Hong Kong	RUS	0	Postgraduate	No
Anson	₹7,475.11	Nov 9, 2020	Hong Kong	RUS	20	Postgraduate	Yes

Feature

Target

Assignment Project Exam Help  
<https://powcoder.com>  
Add WeChat powcoder

# Missing data is a common problem in datasets and needs to be dealt with before applying any ML model

- Missing data refers to the **absence of values** for certain observations and is an unavoidable problem in most data sources
  - e.g. with survey data, some observations may not have been recorded
- **Scikit-learn** does **not** support **missing values** as input, so it is necessary to take one of the following actions
  - remove observations with missing data
  - transform them into permitted values
- The goal of any **imputation** technique is to clean the data to produce a **complete** dataset that can be used to train ML models

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

# Data Type Rectification

<https://powcoder.com>

Add WeChat powcoder

# Data Type Rectification

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- When processing data using Python **dataframes**, some **type mismatch** might occur during the loading of data
- Normal practice is to store
  - Discrete variables as the int type
  - Continuous variables as the float type
  - Categorical variables as the object type
- However, **discrete** variables can sometime be cast (loaded) as **float**
- To correctly identify variable types, both **data types** and **data values** need to be inspected

The "Date" values might be load as "string" type but would be more appropriate to be of "datetime" type

Name	Amount	Date	Issued In	Used In	Age	Education	Fraud?
Daniel	\$2,600.45	1-Jul-2020	HK	HK	22	Secondary	No
Alex	\$2,294.18	1-Oct-2020	HK	RUS	None	Postgraduate	Yes
Adrian	\$1,003.30	3-Oct-2020	HK		25	Graduate	Yes
Vicky	\$8,488.32	4-Oct-2020	JAPAN	HK	64	Graduate	No
Adams	¥20000	7-Oct-2020	AUS	JAP	58	Primary	No
...	...	...	...	...	...	...	...
Jones	₹3,250.11	Nov 1, 2020	HK	RUS	43	Graduate	No
Mary	₹8,156.20	Nov 1, 2020	HK	N/A	27	Graduate	Yes
Max	€7475,11	Nov 8, 2020	UK	GER	32	Primary	No
Peter	₹500.00	Nov 9, 2020	Hong Kong	RUS	0	Postgraduate	No
Anson	₹7,475.11	Nov 9, 2020	Hong Kong	RUS	20	Postgraduate	Yes

Observations

Feature

Target

Assignment Project Exam Help

Missing Value Removal

<https://powcoder.com>

Add WeChat powcoder



# Complete Case Analysis (CCA)

## Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- Discarding those observations where the values in any of the variables are missing
- Can be applied to categorical and numerical variables
- Preserves the distribution of the variables, provided the data is missing at random and only a small proportion of the data is missing
- However, if data is missing across many variables, CCA may lead to the removal of a big portion of the dataset

# Missing value removal involves the removal of an entire observation containing the missing value from the dataset

Name	Amount	Date	Issued In	Used In	Age	Education	Fraud?
Daniel	\$2,600.45	1-Jul-2020	HK	HK	22	Secondary	No
Alex	\$2,294.38	1-Oct-2020	HK	RUS	None	Postgraduate	Yes
Adrian	\$1,003.30	3-Oct-2020	HK		25	Graduate	Yes
Vicky	\$8,488.32	4-Oct-2020	JAPAN	HK	64	Graduate	No
Adams	¥20000	7-Oct-2020	AUS	JAP	58	Primary	No
...	...	...	...	...	...	...	...
Jones	₹3,250.11	Nov 1, 2020	HK	RUS	43	Graduate	No
Mary	₹8,156.20	Nov 1, 2020	HK	N/A	27	Graduate	Yes
Max	€7475,11	Nov 8, 2020	UK	GER	32	Primary	No
Peter	₹500.00	Nov 9, 2020	Hong Kong	RUS	0	Postgraduate	No
Anson	₹7,475.11	Nov 9, 2020	Hong Kong	RUS	20	Postgraduate	Yes

Assignment Project Exam Help

# Data Imputation

<https://powcoder.com>

Add WeChat powcoder

# Imputation

## Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- Imputation is the replacement of missing values with statistical estimates of the missing values
- There are multiple imputation techniques that can be deployed
- The choice of imputation technique will depend on
  - whether the data is missing at random
  - the number of missing values
  - the machine learning model to use

# Imputation techniques vary between numerical variables and categorical variables

## Numerical Variable

- Mean / Median imputation
- Arbitrary number imputation
- End of distribution imputation
- Random sampling imputation
- Missing value indicator augmentation
- Multivariable imputation using chained equations

## Categorical Variable

- Mode imputation
- Random sampling imputation
- Bespoke category imputation
- Missing value indicator augmentation
- Multivariable imputation using chained equations

# Mean or Median Imputation

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- Replacing missing values with variable **mean** or **median**
- Can only be performed in **numerical** variables
- The **mean** / **median** is calculated using a **training dataset** and are used to impute missing data in training, testing, and future datasets
- Use **mean** imputation if variables are **normally distributed** and **median** imputation **otherwise**
- Mean and median imputation **may distort the distribution** of the original variables if there is a **high percentage of missing data**

# Missing values can be replaced with the mean or median of the non-missing values of the feature

Observations

Name	Amount	Date	Issued In	Used In	Age	Education	Fraud?
Daniel	\$2,600.45	1-Jul-2020	HK	HK	22	Secondary	No
Alex	\$2,294.18	1-Oct-2020	HK	RUS	None	Postgraduate	Yes
Adrian	\$1,003.30	3-Oct-2020	HK		25	Graduate	Yes
Vicky	\$8,488.32	4-Oct-2020	JAPAN	HK	64	Graduate	No
Adams	¥20000	7-Oct-2020	AUS	JAP	58	Primary	No
...	...	...	...	...	...	...	...
Jones	₹3,250.11	Nov 1, 2020	HK	RUS	43	Graduate	No
Mary	₹8,156.20	Nov 1, 2020	HK	N/A	27	Graduate	Yes
Max	€7475.11	Nov 8, 2020	UK	GER	32	Primary	No
Peter	₹500.00	Nov 9, 2020	Hong Kong	RUS	0	Postgraduate	No
Anson	₹7,475.11	Nov 9, 2020	Hong Kong	RUS	20	Postgraduate	Yes

mean / median

Feature

Target

The choice of removal or imputation technique is determined by the superiority of model accuracy

	Imputation Technique	# of rows in the training dataset	Accuracy
1	Dropping rows with missing values	392	0.74489
2	Imputing missing values with zero	768	0.7304
3	Imputing missing values with the mean	768	0.7318
4	Imputing missing values with the median	769	0.7357



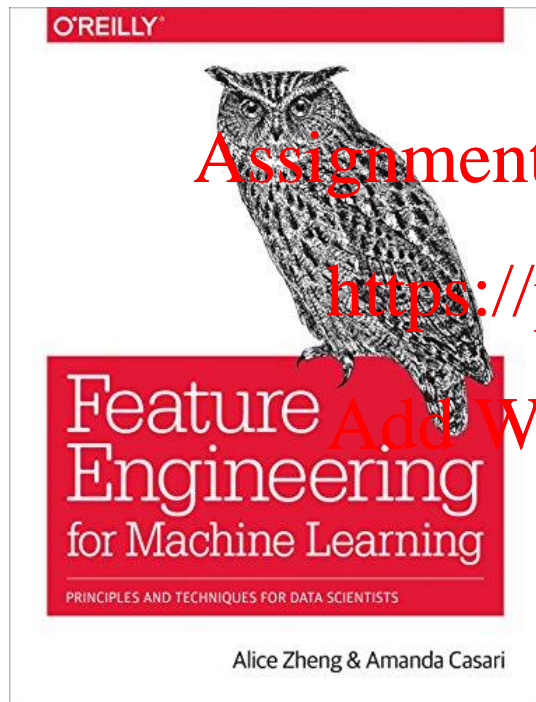
Assignment Project Exam Help

# References

<https://powcoder.com>

Add WeChat powcoder

# References



"Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists"

Alice Zheng & Amanda Casari

O'Reilly Media, April 2018

ISBN-13: 978-1-491-95324-2

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

# Feature Engineering

- "Data Types in Statistics" (<https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/>)
- "Types of Data & Measurement Scales: Nominal, Ordinal, Interval and Ratio" (<https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/>)
- "Measures of Central Tendency" (<https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php>)
- "Scales of Measurement and Presentation of Statistical Data" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6206790/>)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

# Financial Datasets

- “yfinance 0.1.54”, 2019  
<https://pypi.org/project/yfinance/>
- "quandl/quandl-python", 2019  
<https://github.com/quandl/quandl-python>

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

# Public Datasets

- Academic Torrents (<https://academictorrents.com/browse.php?cat=6>)
- Awesome Public Datasets (<https://github.com/awesomedata/awesome-public-datasets>)
- Awesome JSON Datasets (<https://github.com/jdorman/awesome-json-datasets>)
- Common Crawl (<http://commoncrawl.org/the-data/>)
- DataHub Datasets (<https://datahub.io/search>)
- Kaggle Datasets (<https://www.kaggle.com/datasets>)
- GitHub Archive (<http://www.gharchive.org/>)
- GitHub COCO-Stuff Datasets (<https://github.com/nightr0me/cocostuff>)
- Harvard Resources for COVID-19 (<https://dataverse.harvard.edu/dataverse/2019ncov>)
- GitHub COVID-19 Data (<https://github.com/owid/covid-19-data/tree/master/public/data/>)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

# Public Datasets

- Coronavirus Source Data (<https://ourworldindata.org/coronavirus-source-data>)
- OCHA Novel Coronavirus (COVID-19) Cases Data (<https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>)
- World Bank Open Data (<https://data.worldbank.org/>)
- Hong Kong Government Open Data (<https://data.gov.hk/>)
- US Government Open Data (<https://www.data.gov/open-gov/>)
- Taiwan Government Open Data (<https://data.gov.tw/>)
- Dataquest – 18 Places to Find Free Data Sets for Data Science Projects (<https://www.dataquest.io/blog/free-datasets-for-projects/>)
- Google Dataset Search (<https://datasetsearch.research.google.com/>)
- UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets.php>)
- KDnuggets Datasets for Data Mining and Data Science (<https://www.kdnuggets.com/datasets/index.html>)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

# Public Datasets

- Google BigQuery Public Datasets (<https://cloud.google.com/bigquery/public-data/>)
- Google Research Datasets (<https://research.google/tools/datasets/>)
- Microsoft Public Data Sets for Testing and Prototyping (<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-public-data-sets>)
- Amazon Web Services Registry of Open Data (<https://registry.opendata.aws/>)
- Pathmind Open Datasets (<https://pathmind.com/wiki/open-datasets>)
- Lionbridge 15 Best Audio Datasets for Machine Learning (<https://lionbridge.ai/datasets/12-best-audio-datasets-for-machine-learning/>)
- Google COVID-19 Public Datasets (<https://console.cloud.google.com/marketplace/details/bigquery-public-datasets/covid19-dataset-list?preview=bigquery-public-datasets>)
- Google Public Data ([https://www.google.com/publicdata/directory?hl=en\\_US&dl=en\\_US#!](https://www.google.com/publicdata/directory?hl=en_US&dl=en_US#!))
- Tableau - Coronavirus (COVID-19) Global Data Tracker (<https://www.tableau.com/covid-19-coronavirus-data-resources>)

Assignment Project Exam Help

<https://powcoder.com>  
**THANK YOU**

Add WeChat powcoder