



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

SC1015 : Review Lecture

Exploratory Data Analysis

Dr Smitha K G





SC1015

Admin Announcements

1. Detailed solutions to the Lab Exercises will be posted every week after the Lab Week is over.
2. Mini-Project details will be posted by Week 6.

LAMS Completion Status

Module 1 Part 1 : Above 940 – Quiz solutions posted
 Module 1 Part 2 : Above 850 – Quiz solutions posted
 Module 2 Part 1 : Above 750– Quiz solutions posted (check)
 Module 2 Part 2 : Above 550 – Complete by Exercise 3
 Module 3 : Above 200 – Complete by Exercise 4 (W6)

Graded Lab Exercises in Weeks 4, 6, 7 – 5% each. DS Theory Quiz in Recess Week : 8 March, Friday.

Attendance Mandatory

Understand the data

- Numerical variables
- Categorical variables

Some categorical can be treated as numerical and visa versa

Clean the data

- Remove null values
- Outliers can be excluded

Remove noise (unwanted materials)

Analysis of relationship between variables

- Using different kinds of plots
- Univariate, bivariate and multivariate

Pairplot, scatterplot, swarmplot, relyplot, Heatmap

Getting Started Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 6,330 teams · Ongoing

Overview Data Notebooks Discussion Leaderboard Datasets Rules [Join Competition](#)

Overview

Description

Evaluation


Tutorials

Frequently Asked Questions

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description



Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Practice Skills

- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting

Acknowledgments

The [Ames Housing dataset](#) was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

Notebooks > 873 discussion topics >

Data Science

Exploratory Data Analysis

Target : Predicting SalePrice

Understand the variable **SalePrice**

Understand **all** the other Variables

Understand all **mutual** relationships

Clean and prepare the Training Data

Then think of creating your **Model** ;-)

What about other “interesting” questions on the same dataset?

<https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python>



Questions you have for Ex3...

1. Does **having a Garage or not having a Garage** have “any” effect on **SalePrice**?

We have 6 variables for **Garage** ('GarageType', 'GarageYrBlt', 'GarageFinish', 'GarageCars', 'GarageArea', 'GarageQual', 'GarageCond')

Which of them will have the best effect on the SalePrice of a house within this dataset?

2. Does the **SalePrice** of a house get affected by how **recently** it got **Remodeled**?

How would you justify your answer in terms of data?

Read data description carefully, and then start coding.



Questions you may ask ...

Start thinking of “special/ specific” questions next.

1. Which of the houses sold for the **maximum SalePrice**?
2. **Why** do you think that specific house commanded the price?

How would you justify your answer in terms of data?

Code a little to find the target house, and then analyze.



Questions you may ask ...

Finally, think of “extension” questions!

What will be your strategy to **buy and remodel** a house if you plan to **flip houses** and make **maximum profit**?

How would you justify your answer in terms of data?

Think about it; take your time to formulate the problem.

Airbnb in Seattle — Data Analysis

Understand Airbnb rental landscape in Seattle



For all prospective Airbnb hosts in Seattle, I will answer these questions in this article:

- when to rent to maximise revenue?
- when is the off-peak season for maintenance?
- common group size of Seattle travellers, is it 2 or family or 4 or larger?
- bedroom configurations to maximise booking rates?
- how to achieve a good rating?
- do hosts with higher rating have higher revenue?
- amenities to include?

Data Science

Showcase : Khok Hong Jing

You may build a similar portfolio yourself.

<https://github.com/jinglescode>

<https://jinglescode.github.io/>

If you do not have a GitHub profile, create one now!

If you want more of a competition, get into Kaggle.

Showcase your skills in Coding as well as in Analysis.

<https://jinglescode.github.io/2019/07/13/airbnb-in-seattle-data-analysis/>



Questions you have asked ...

Year and Month : Numeric or Categorical?

YearBuilt : Categorical?

YrSold : Categorical?

$\text{YrSold} - \text{YearBuilt} = \text{Age!}$

Think of creating new features out of the given ones.

This, done smartly, is also called Feature Engineering. 😊



Questions you have asked ...

OverallQual : Numeric or Categorical?

Numeric : Statistics Work

Category : Levels Different

May be “Ordinal” Category

Think if you can use the same feature in both ways.

This, done smartly, will help you with diverse inferences.

Numeric vs Numeric

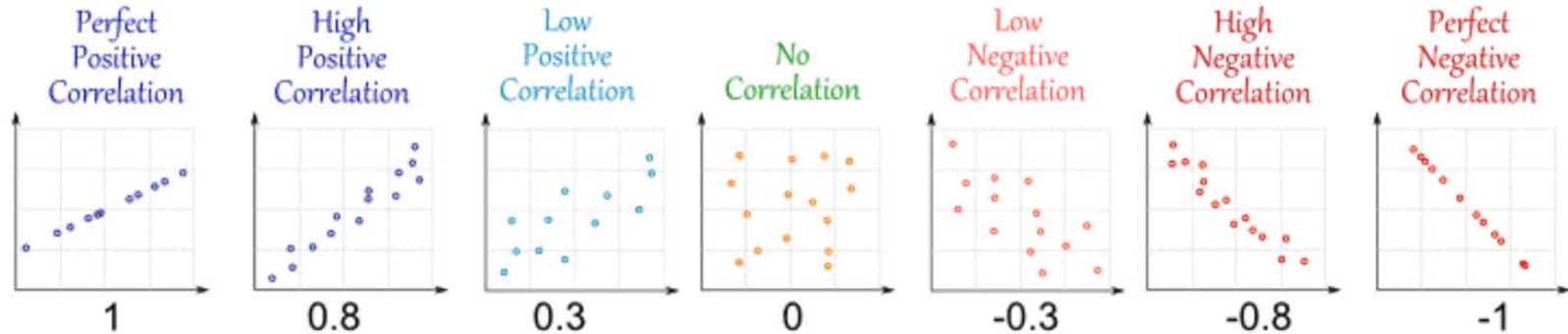
Numeric vs Ordered Categorical

Numeric vs Unordered Categorical

Ordered Categorical vs Ordered Categorical

Unordered Categorical vs Unordered Categorical

Relationship Complicated!



What do you want to explore visually? Check this : <https://www.data-to-viz.com/>

seaborn

Know your plots!

Uni-Variate
Numeric

boxplot
histplot
violinplot

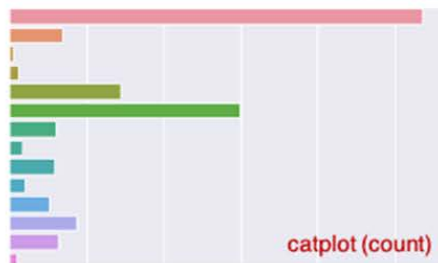
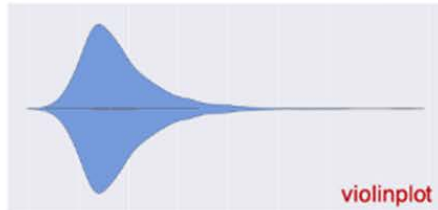
Uni-Variate
Categorical

catplot

Bi-Variate
Mixed

jointplot
boxplot
heatmap
(groupby)

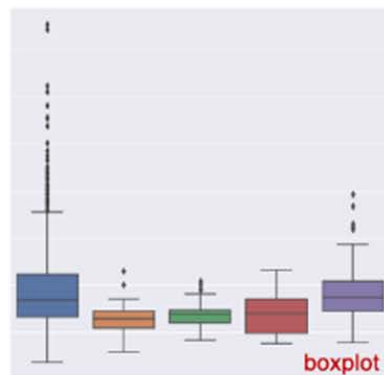
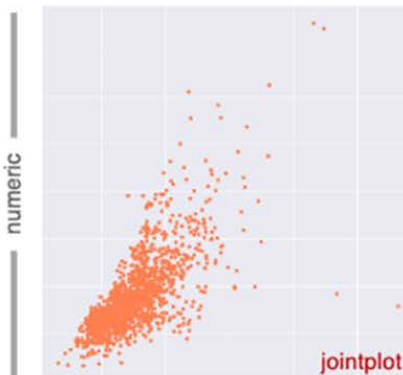
Uni-Variate Plots



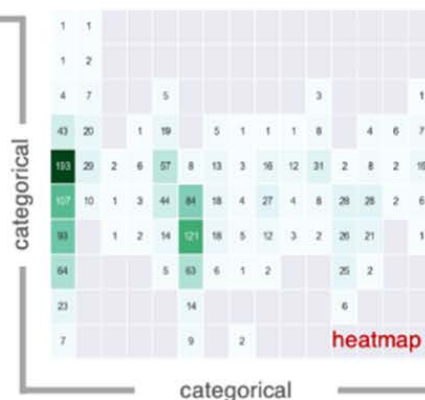
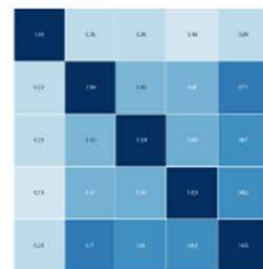
numeric

categorical

Bi-Variate Plots



heatmap for correlations



Getting Started Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 6,330 teams · Ongoing

Overview Data Notebooks Discussion Leaderboard Datasets Rules [Join Competition](#)


Overview

Description

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description



Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Practice Skills

- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting

Acknowledgments

The [Ames Housing dataset](#) was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

Notebooks > 873 discussion topics >

Data Science Machine Learning

Prediction : Numeric Regression

Model : $\text{SalePrice} = f(\text{Variables})$

Given	Some Houses as Train Data
Learn	The Formula for SalePrice
Predict	Estimate SalePrice for Test

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>