Proud Cockroaches

COVID-19 DATA ANALYTICS & PREDICTION MODEL

Discrete Structure (CO1007)

MAIN TASKS

01 Data collection

Collect data on the disease situations around the world.

02 Data analysis

Provide analysis and developments.

03 Model prediction

Predict the epidemic situation in the next 1 week - 3 months

04 Model analysis

Testing and stuffs (?)

01

DATA COLLECTION

Collect data and provide analysis.

DATA ANALYTICS (ALL TASKS)

Có vài bước chính trong "analyzing" a.k.a. **Data analytics**:

Bước 0: Collect data.
Bước 1: **Exploratory data analysis (EDA)**.
Còn gọi là "Data profiling", mục đích là để cho team hiểu hơn về cái dataset này, đào sâu được càng tốt. Các task cụ thể là:

- **Data cleaning**: check duplicate rows, check NA values để xoá hoặc fill mấy ô data trống, check mấy dữ liệu bị ghi sai,...
- **Statistical analysis**: cái này thống kê chung chung, gồm:
  - Data có bao nhiêu hàng, bao nhiêu cột? Liệt kê mỗi cột (có thể gọi là variable/attribute) chứa data gì (short description + data type)
  - Bảng câu hỏi mà mình cần làm
  - **Descriptive analysis**: Trong code thường sẽ có 1 cái bảng thống kê cho dataset, bảng đó để thống kê các **numerical variables** hoặc **categorical variables**, (ví dụ ở 2 slide sau)
- **Visualization**: thực ra cái visualization này cũng để hỗ trợ cho bước ở trên, cũng là descriptive analysis (quantitative and qualitative analysis) (slide 7).

Bước 2: Model fitting
Bước 3: Model evaluation
Bước 0 và 1 là task 1 và 2, bước 2 và 3 là task 3 và 4.

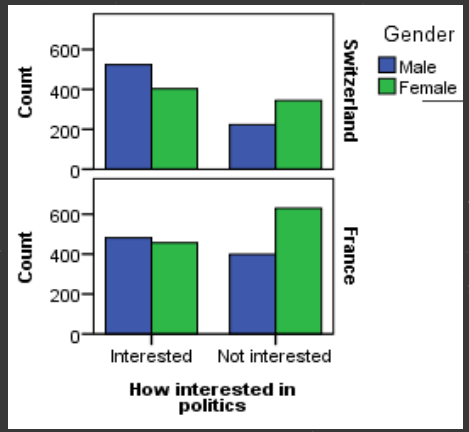
	Name	Team	Number	Position	Age	Height	Weight	College	Salary
count	364	364	364.000000	364	364	364	364.000000	364	3.640000e+02
unique	364	30	NaN	5	22	17	NaN	115	NaN
top	Cleanthony Early	New Orleans Pelicans	NaN	SG	24.0	6-9	NaN	Kentucky	NaN
freq	1	16	NaN	87	41	49	NaN	22	NaN
mean	NaN	NaN	16.829670	NaN	NaN	NaN	219.785714	NaN	4.620311e+06
std	NaN	NaN	14.994162	NaN	NaN	NaN	24.793099	NaN	5.119716e+06
min	NaN	NaN	0.000000	NaN	NaN	NaN	161.000000	NaN	5.572200e+04
20%	NaN	NaN	4.000000	NaN	NaN	NaN	195.000000	NaN	9.472760e+05
40%	NaN	NaN	9.000000	NaN	NaN	NaN	212.000000	NaN	1.638754e+06
50%	NaN	NaN	12.000000	NaN	NaN	NaN	220.000000	NaN	2.515440e+06
60%	NaN	NaN	17.000000	NaN	NaN	NaN	228.000000	NaN	3.429934e+06
80%	NaN	NaN	30.000000	NaN	NaN	NaN	242.400000	NaN	7.838202e+06
max	NaN	NaN	99.000000	NaN	NaN	NaN	279.000000	NaN	2.287500e+07

NUMERICAL VARIABLES

CATEGORICAL VARIABLES

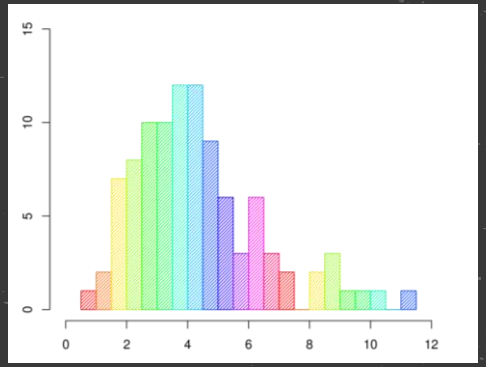
```
df['Product Division'].value_counts()

Footwear        266742
Apparel         248699
Accessories      51046
Name: Product Division, dtype: int64
```

Có thể lập bảng **count** (ví dụ code Python)



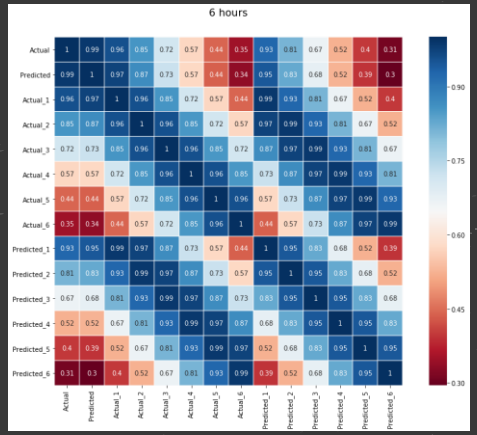
Hoặc có thể tạo 1 cái count barchart cho số liệu loại categorical này.

DATA VISUALIZATION



Graphings

Lập các histogram, line plot, count plot, boxplot v.v.. cho vài variables



Correlation matrix

(Google :v)

Custom visuals

VD: Tạo 1 graph về số người chết do bệnh nền hoặc ko, hoặc 1 count barplot cho các nhóm nhiễm chủng Covid khác nhau....

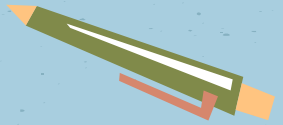
DISTRIBUTION NOTES

Cần xác định cái cột **variable** nào mình muốn xét distribution. Đầu tiên tạo 1 cái visual (histogram chẳng hạn). Sau đó chọn 1 loại **distribution** muốn xét dựa vào **tính chất** cái variable của mình: Normal, Poission, Gamma, Exponential... (mỗi cái có công thức riêng, cần xem kỹ). Cứ visualize xong rồi **so sánh** hình dạng. VD: theo normal distribution, nếu cái shape nó không được "normal" cho lắm thì có thể kết luận nó là không normal, hay right skewed, hay left skewed gì đó... (?)

02

PREDICTION MODEL

Use various types of models to predict the situation in the next time interval.

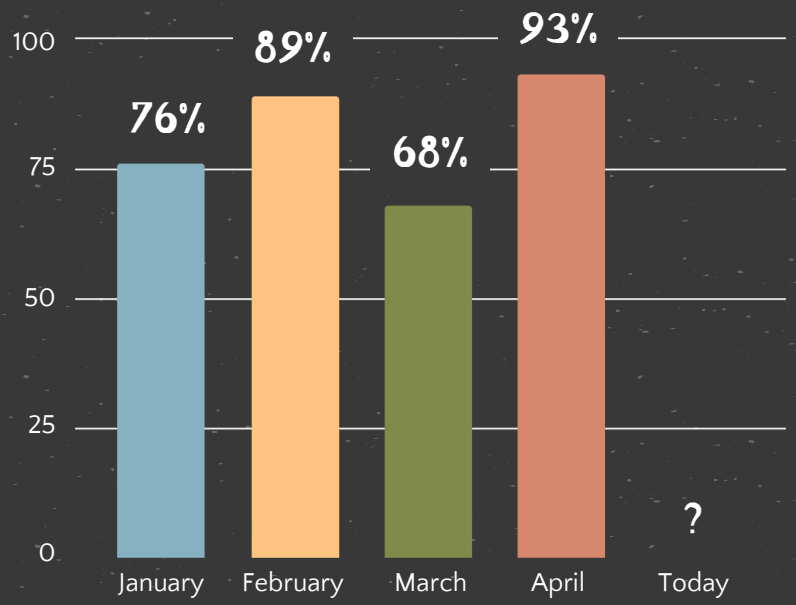


BUILDING MODEL (TASK 3)

Bước 1: Đầu tiên phải chọn những input variables nào mình cần đưa vào model (bất kỳ model nào cũng dc, nên làm model đơn giản quen thuộc trước).

- **Explanatory variables X** (thuật ngữ có thể gọi là features/predictors): là những variables mình cho rằng nó sẽ ảnh hưởng tác động đến target
- **Responsive variable Y** (label / target): Mục tiêu mình cần dự đoán là gì?
  - Nếu variable đó là categorical variable -> chọn **classification models** (logistic regression, random forest classification,...)
  - Nếu variable đó là numerical variable -> chọn **regression models** (linear regression, random forest regression,...)

Bước 2: Chia cái dataset vừa chọn ra làm train set và test set, ratio Train:Test là 8:2 hoặc 7:3 hoặc 75:25 tùy :v Chia xong sẽ có: X_train, y_train, X_test, y_test. Dùng Train set để bỏ vào chạy training/fitting model. Dùng Test set để làm Task 4.



MODEL ANALYSIS (TASK 4)

#Về testing và model analysis

Sau khi dựng model, model sẽ **predict** cho mình cái target mà mình muốn predict. VD: nhét đống x_test vào model fittedmodel.predict(x_test) => nó sẽ cho ra 1 dãy prediction, gọi là **y_predicted**, tương ứng với số hàng của x_test mình bỏ vào,

Vậy là mình sẽ có **y_test** (hay **y_actual**) và **y_predicted**. Dựa vào 2 cái này, mình sẽ tính ra các **model evaluation metrics**, nhằm để đánh giá **model accuracy** xem nó predict tốt hay ko.

Cá nhân tui sẽ chia ra 2 loại model **Classification** và **Regression**, mỗi loại có cách đánh giá riêng (các slides kế tiếp).

confusion matrix

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

```
Confusion Matrix:
 [[98037 39072]
 [ 2159  2221]]

Classification report:
              precision    recall  f1-score   support

           0       0.98      0.72      0.83    137109
           1       0.05      0.51      0.10      4380

    accuracy                           0.71    141489
   macro avg       0.52      0.61      0.46    141489
weighted avg       0.95      0.71      0.80    141489
```

CLASSIFICATION MODELS

REGRESSION MODELS

Có vài **evaluation metrics** mình cần quan tâm:

- **R-squared** (R^2 / coefficient of determination): nên chỉ nêu lên rồi giải thích ý nghĩa của nó thôi.
- **MSE**, **RMSE**, **MAE**, **MAPE**, dựa trên y_actual và y_predicted. **Error metrics** càng nhỏ thì model đó càng có **accuracy** cao.

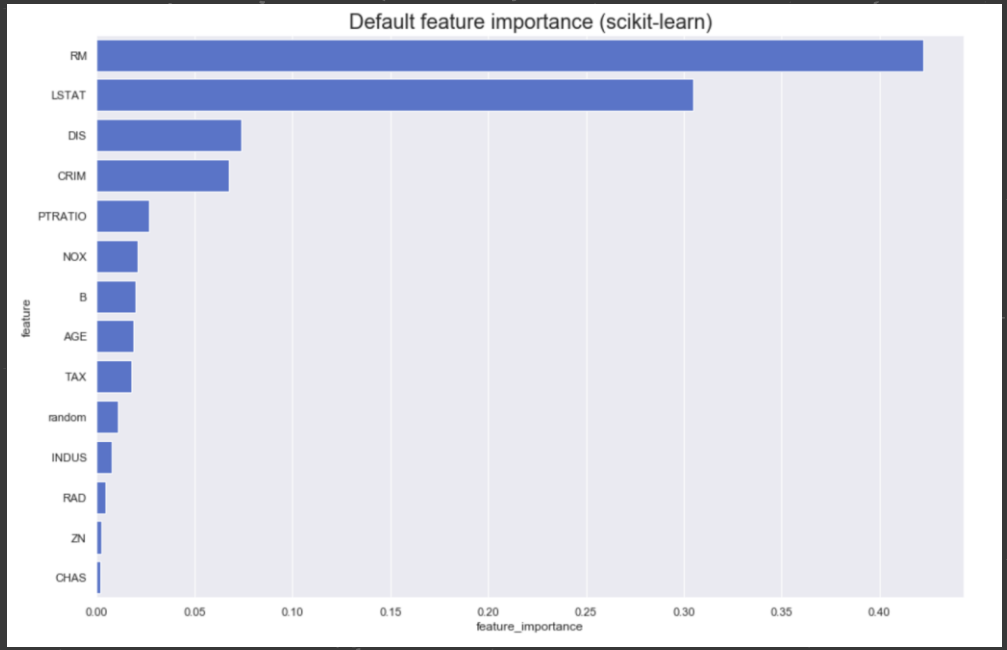


MODEL ANALYSIS (TASK 4)

#Về influencing factors

- Đối với **classification**, ta sẽ có cách tính **feature_importance**. Để giải thích thì mình xem "top những explanatory variables nào điểm cao nhất thì nó ảnh hưởng tới outcomes của model nhiều nhất". (?)

VD: feature importance của random forest (?):



MODEL ANALYSIS (TASK 4)

#Về influencing factors (tiếp)
Đối với **regression**, trong linear regression có mấy cái coefficient với p-value, chọn cái nào [cao nhất (?)] thì nó là most influencing factors.

Lặp lại những bước building và testing model ở trên để làm ra 2-3 cái model, và dựa vào những **evaluation metrics** (của từng loại classfication hay regression), thì cuối cùng mình sẽ đem ra so sánh xem thử cái nào có **accuracy** cao hơn thì cái đó best.

time series analysis

THANKS LOL