

DS049-Exploratory Data Analysis (EDA) 2

أكاديمية طويق
Tuwaiq Academy



By: Nourah Almutlaq

مقدمة

تخيل معي خلال رحلتك للبحث عن الفلم المناسب لك لنهاية الاسبوع، قد تجد نفسك أمام عدة خيارات مختلفة ولا بد لك من تقييم كل خيار على حده. لنفرض أنك وجدت الفلم X، لنقيّم هذا الفلم سوف تبحث عن نوع الفلم ومن ثم نسبة التقييم وطاقم التمثيل والمخرجين لهذا الفلم، إذا أعجبك سوف تشاهد الإعلان الترويجي وبعد كل هذه الخطوات يأتي اتخاذ القرار بالمشاهدة أو البحث عن فلم آخر مجدداً.

تحليل البيانات الاستكشافي (EDA - Exploratory Data Analysis)

بنفس المنهجية الموضحة سابقاً فإن تحليل البيانات الاستكشافي ماهو إلا مرحلة تتيح لك التعرف على بياناتك واستكشافها، وتحديد أهمية كل مُدخل بما يتناسب مع أهدافك للوصول إلى رؤية واضحة.

- مرحلة EDA تم تطويرها من قبل عالم الرياضيات الأمريكي John Tukey في السبعينات.
- هي منهجية يستخدمها علماء البيانات بهدف تحليل مجموعات البيانات واستخلاص خصائصها.
- تُساعد في تحديد أفضل الطرق للتعامل مع البيانات للحصول على الإجابات التي نحتاجها.
- تُسهل على علماء البيانات اكتشاف الأنماط والحالات الشاذة واختبار الفرضيات أو التحقق منها.
- خلال هذه المرحلة، غالباً ما يستخدم أحد الطرق التالية:

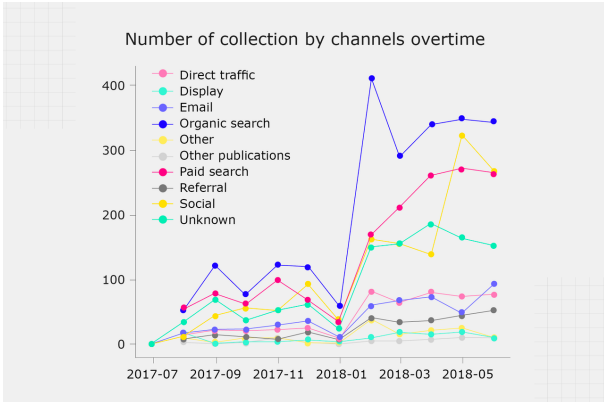
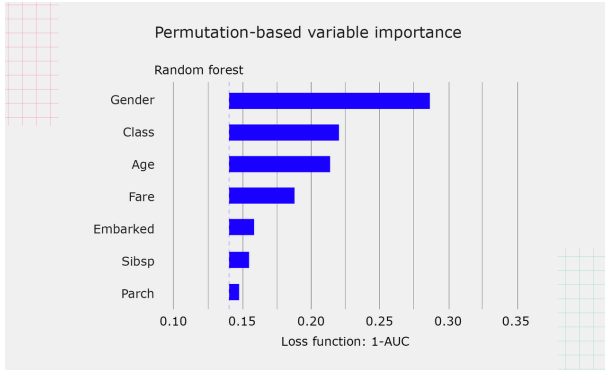
- الأساليب الإحصائية مثل: t-test.
- عرض البيانات (data visualization methods)، مثل: bar plot.

- توفر لغة Python عدة مكتبات تُسهل استكشاف والتعرف على البيانات منها: NumPy, Pandas, Matplotlib.

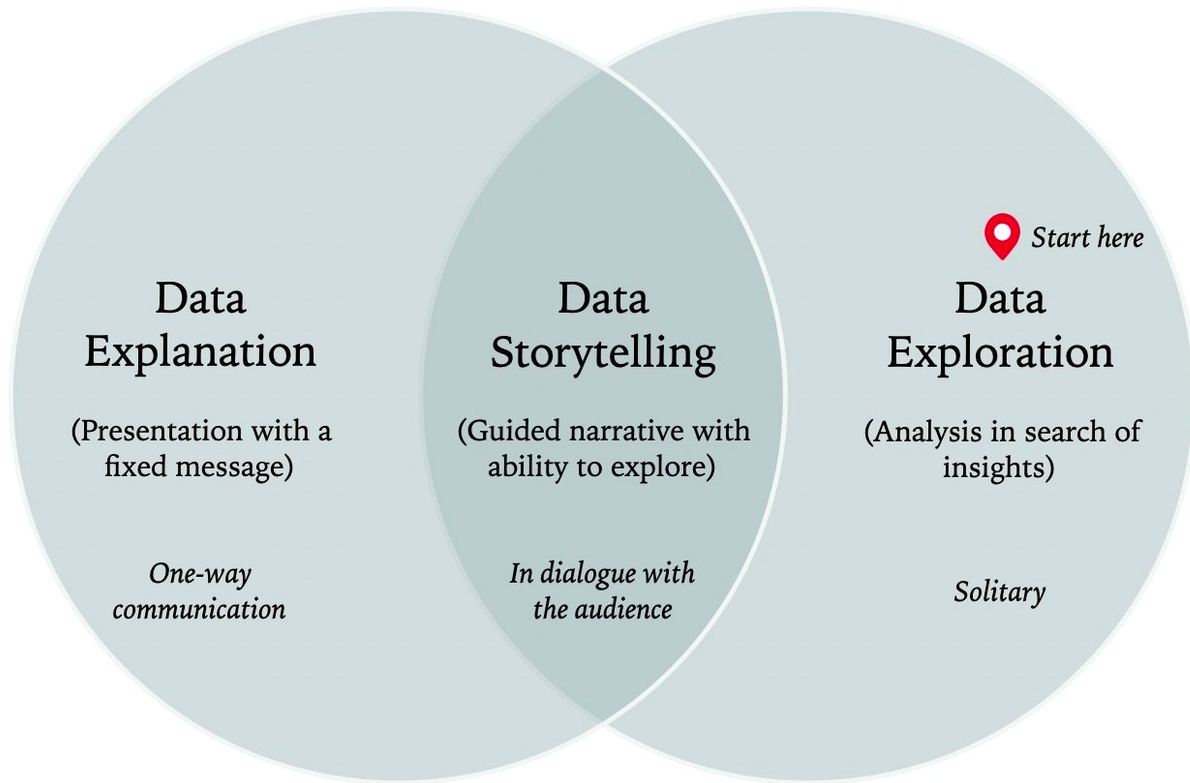
أهمية مرحلة EDA

- للمساعدة في أخذ نظرة عامة عن البيانات قبل إجراء أي افتراضات.
- اكتشاف الأخطاء و فهم الأنماط داخل البيانات.
- اكتشاف القيم المتطرفة (outliers) وإيجاد العلاقات المثيرة للاهتمام بين المتغيرات.

الفرق بين Exploratory و Explanatory Visualization

Exploratory (Data Exploration)	Explanatory (Data Explanation)
يهدف إلى ربط مصادر البيانات، وتحديد العلاقات داخل البيانات، وفهم المقاييس.	يهدف إلى جمع insights وربطها بشكل منطقي ثم محاولة فهم الجمهور ومايهمهم من هذه الرؤى وأخيرا عرضها عليهم.
يُستخدم لعرض البيانات على الخبراء، الذين لديهم معرفة مسبقة بالموضوع.	يُستخدم لعرض البيانات على الناس غير الخبراء الذين ليس لديهم معرفة مسبقة بالموضوع.
يُمثل البيانات باستخدام رسومات معقدة.	يُمثل البيانات باستخدام رسومات مفهومة.
تُستخدم لأغراض تحليلية.	لا تُستخدم لأغراض تحليلية ولكن بغرض تفسير المعلومات.
يمثل نقطة البداية للعمل مع البيانات حيث لا يمكن الانتقال إلى شرح النتائج للآخرين بدون استكشاف البيانات وفهمها.	يمثل الخطوة الأخيرة للعمل مع البيانات حيث تساعد الأشخاص على اتخاذ القرارات.
	
Click to enlarge	Click to enlarge

خلال الانتقال بين الخطوتين السابقتين نمر بمرحلة تسمى: Interactive Data Storytelling.



خطوات تحليل البيانات الاستكشافي

- استكشاف بياناتك من حيث العدد والأعمدة والصفوف وأنواع البيانات الموجودة عن طريق أحد الأوامر التالية:

```
# Overview of the dataset
data

# Print the first five records in the dataset
data.head()

# Print the last five records in the dataset
data.tail()

# Print numbers of columns & rows
data.shape
```

- استكشاف القيم المفقودة ونوع البيانات لكل عمود.

```
# Print the number of missing values and data type for each column
data.info()
```

بعد ذلك سوف ننتقل لخطوات ذات تفاصيل ومعلومات أكثر عن طريق استخدام الأساليب الإحصائية والرسوم البيانية.

أولاً: التحليل الوصفي (Descriptive Analysis)

عبارة عن وصف وتلخيص للبيانات من خلال:

- ملخص إحصائي رقمي (Numerical Statistical Summary)

```
# Print some main statistical calculations for the numerical columns
data.describe()
```

- ملخص إحصائي فئوي (Categorical Statistical Summary)

```
# Print some main statistical calculations for the Categorical columns
data.describe(exclude='number')
```

بعد مرحلة فهم البيانات إحصائياً، سوف ننتقل لتوضيح العلاقة بين المتغيرات عن طريق رسوم بيانية تساعد على استنتاج الحقائق.

ثانياً: التحليل أحادي المتغير (Univariate Analysis)

يهدف إلى تقديم ملخصات إحصائيات لمتغير واحد أو عمود واحد في مجموعة البيانات عن طريق رسوم بيانية تختلف بحسب المتغير إذا كان Numerical أو Categorical.

Numerical Data	Categorical Data
Histogram	Bar Chart (Ordinal)
Density	Pie Chart (Nominal)

ثالثاً: التحليل متعدد المتغيرات (Multivariate Analysis)

يُستخدم لفهم العلاقة بين متغيرين (Bivariate Analysis) أو أكثر (Multivariate Analysis) عن طريق رسوم بيانية.

- العلاقة بين متغيرين (Bivariate Analysis)

--	--	--

Numerical Vs. Numerical	Numerical Vs. Categorical	Categorical Vs. Categorical
Scatter Plot	Bar Plot	Clustered Histogram
Line Chart	Scatter Chart	Clustered Bar (Side-by-side Bar)
	Violin Plot	
	Box plot	
	Clustered Bar	
	Swarm Plot	

• العلاقة بين أكثر من متغير (Multivariate Analysis).

Multivariate Analysis
Bar Chart
Scatter Chart
Line Chart

من الطرق الأخرى:

Univariate Analysis

Distribution Plots	Comparison Plots	Composition Plots
Histogram	Bar Chart	Pie
Frequency Polygon Plot	Line Chart	Waffle

Density Plot	Run Chart	Tree Map
Box Plot	Sparkline	Waterfall
Violin Plot	Lag Plot	
Strip Plot	Circular Area (Summary) Plot	
Swarm Plot	Cartograms plot	

Multivariate Analysis

Distribution Plots	Comparison Plots	Composition Plots	Relationship Plots
Scatter Plot	Side-by-Side Bar Chart	Stacked Area	Scatter Plot
Joint Plot	Heatmap	Stacked Bars	Bubble Chart
Sterogram Plo	Run Chart	Tornado	
Surface Area	Cartograms plot		
Level Curve			
Box Plot			
Violin Plot			

Resources:

- <https://app.pluralsight.com/library/courses/exploratory-data-analysis-python/table-of-contents>
- <https://www.ibm.com/topics/exploratory-data-analysis>
- <https://www.futurelearn.com/info/courses/data-visualisation-with-python-matplotlib-and-visual-analysis-sc/0/steps/314426>