

Exploring Data and knowledge
discovery

د.أبي صندوق

2020\7\1

RB Informatcs; محتوى مجاني غير مخصص للبيع التجاري

data engineering

Data Engineering

Data in ISE:

تتعامل نظم المعلومات مع المعلومات والمستندات (معلومة مفيدة للصراحة)، وآلية القيادة لكل نظم المعلومات هي ال data حيث يعتمد النجاح في هندسة نظم المعلومات على النجاح في هندسة ال data

تستخدم ال data الخام في كل المجالات عند تشغيل نظام متضمنة المناقلات اليومية، تنميط العملاء costumer profiling , تحليل الاستراتيجيات , إدارة أداء الأعمال , صنع القرار والتحليلات التنبؤية والوصفية .

المعلومات المقادة بال data أصبحت الآن محتومة وخوارزميات التعامل مع ال data تدخل بالأعمال بشكل كبير وأصبحت 80٪ من الشركات تهتم بموضوع ال DE بشكل أساسي وقامت بتهيئة تطوير الكفاءة المعتمدة في مجال محو الأمية للبيانات وتعترف بالنقص الشديد بها حيث يحتاج كل ذلك الى ثقافة التعامل مع ال data

والتركيز من خلال ال machine learning سوف يقوم بالإزاحة من الخوارزمية لل data ذات القيمة العالية

ما هي نظم البيانات Data System وما الذي يمكن عمله بها؟

Current Climate

المعلومات موجودة في كل مكان، شركات، أعمال، حكومات، مدارس..... ويتم الاحتفاظ بها في مستندات documents يقوم عليها تسير الأعمال hard copies or /and soft copies ويجب أن تتم صيانتها

Business flow, legal reasons, archive reasons, competitive reasons, decision support, etc...

والمعلومات يتم خلقها وتخزينها ومعالجتها وصيانتها وإهمالها لتوجد معنى لتواجد شركة ما حيث يتواجد هذا المعنى عن طريق قدرتها على التحكم بالمعلومات

ووصول ونشر واستعمال المعلومات هذه الأيام ينافي فهم البشر.

Data

- جمع ال objects من ال attributes الخاصة بها
- بناء blocks من المعلومات
- يمكن أن يعبر عن كل أنماط المعلومات عندما يتم وضعها بالسياق المناسب
- مع عمر تقني فإن تخزينها وصيانتها أمر بسيط

Data - Examples

- Relational Databases
 - Logical model: ERD
 - Physical model: tables, indices
 - Query language: SQL
- Not representing the real world.
- Only one type of data: the relation.
 - عبء التحويل Conversion burden
- Homogeneous data. البيانات المتجانسة
- Other RDBMs limitations: changes to schema, short-lived transaction, etc...

Data Types

- يوجد ثلاثة أنماط مهمة للتعامل مع ال data:

Structured data:

الأسهل في التعامل وتكون غالبا في قواعد البيانات

Semi-structured data:

تكون غالبا في ملفات ال XML

Unstructured data:

الأكثر انتشارا، عشوائية غالبا مثل الملفات (صوت وغيره).
حيث أن الداتا موجودة في كل مكان ومنتشرة بشكل أكثر بكثير من المستندات
ويجب أن تتم صيانة الداتا المخزنة من أجل تدفق الأعمال المناسب والملائم ويجب الاحتفاظ بها كمرجع مستقبلي
ولكن ما الذي يجب الاحتفاظ به؟ كل شيء؟
ممكن أن نصل لنقطة لا يعود ممكنا فيها التخزين أكثر، على الأقل فيزيائيا وهنا يجب اتخاذ قرار بالداتا التي يجب
الاحتفاظ بها والتي من الممكن خسارتها

Data Flood

تضاعف الداتا في العالم كل سنتين (Moore's Law).

تم تخمين حجم الداتا العالمية عام 2015 رب 8 Zetabyte.

تبعاً للحاسب الضخم IBM فإن 2.5 Exabyte من الداتا يتم توليدها كل يوم في عام 2012 وهذا الرقم كبير في معايير أي شخص.

75٪ من الداتا تكون من نوع unstructured قادمة من مصادر مثل .text,voice,video.

وبالإشارة الى amazon نحن هنا نتكلم عن 50 billion إلى 100 billion قطعة من المعلومات في اليوم (عام 2016).

Data – Digital Waste

الداتا المهمة وغير المهمة توضع مع بعضها وقسم منها لا يمكن الاستفادة منه (غير مفيد او لا يمكن استرجاعه)

استرجاع الداتا أصبح مرهقا بوجود الكم الهائل من الداتا عديمة النفع التي يتم البحث عنها

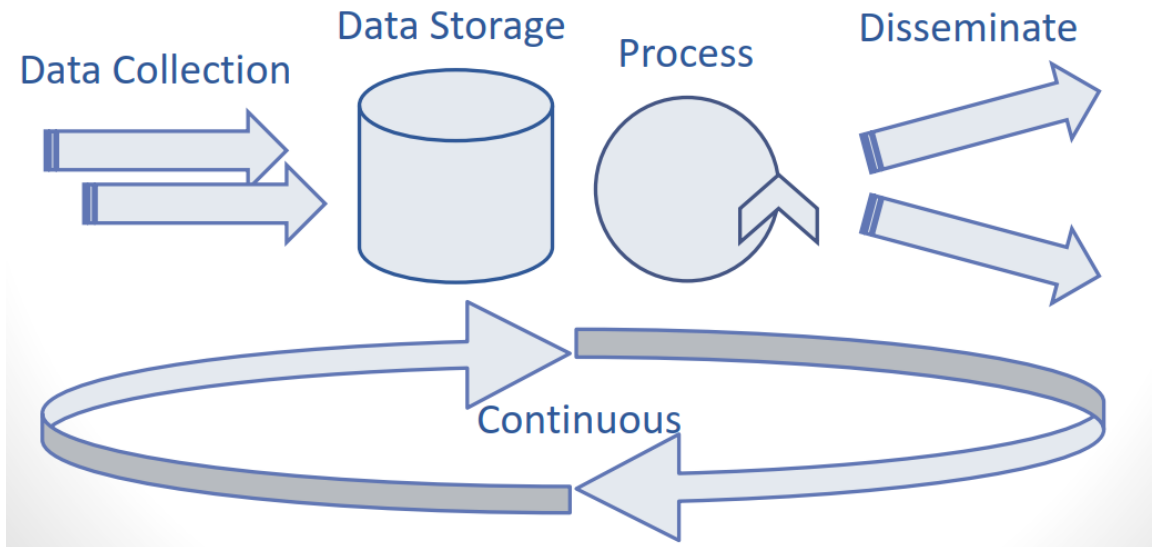
من حيث المبدأ، لا يفترض حذف الداتا ويجب الحفاظ عليه من التخريب إذا كان من الممكن الاستفادة منها ولكم عند

الوصول للـ Data flood يتم التخلص منها من الأقل أهمية الى الأكثر أهمية

وممكن ممانعة عملية الحذف بإعطائها لـ system يستفيد منها

ما الذي يجب القيام به لتجنب الوقوع في الـ digital-waste؟

علينا بدراسة دورة حياة الداتا، من اين تأتي، كيف يتم استرجاعها؟ الى اين تذهب بعد ذلك؟



Data lifecycle _ Aspects

الهدف من إدارة دورة حياة الداتا DLM هو تحديد الأولويات ومعالجة الداتا حيث لا تكون كل الداتا مثل بعضها ولا

يمكن أن تتواجد في متناول اليد في نفس اللحظة ولكنها تشترك بدورة حياة معينة.

Document Retention :

الاحتفاظ بالمستندات يعتمد على تعريف أهمية الداتا والتخطيط لصيانتها.

الحالة الشائعة:

الداتا يجب أن يتم اهمالها ما لم تكن business جيد او هناك سبب قانوني للاحتفاظ بها
أمن المعلومات:

- حماية البيانات هو شيء أساسي في أي منظمة لاستمرارها بالمصادقة على التسجيلات الخاصة وإمكانية استرجاع الداتا

Distribution Control:

خاصة بالعمر الرقمي

Data Lifecycle _ Collection

طرق جمع الداتا عديدة ولا يوجد طريقة متأصلة أكثر من الأخرى ولكن الشيء الأهم توثيق عملية الجمع للاستفادة منها لاحقاً ويجب أن تكون حاسماً حول ما يجب جمعه

مصادر الداتا الأولية تتضمن:

Interviews, Experimentation, Surveys, User shares own experiences, User photos, media, Gadgets info, Statistical analysis of samples, etc...

مصادر الداتا الثانوية تتضمن:

Reviews, Perceived life-style, Inferred tags/labels, etc...

ممارسة ال minimalism على مستوى المنظمة:

لتصغير حجم الداتا المخزنة قدر الإمكان الحل الأفضل هو عدم ادخال داتا من الأساس اذا يجب تحديد اذا منا بحاجة الى الداتا التي نقوم بجمعها وعلى مستوى المنظمة يجب جمع الداتا التي تفيد كل قسم في المنظمة واهمال أي شيء اخر لا يفيد أي قسم

اذا لم نستطع اتخاذ قرار بناء على الداتا الموجودة، لا يمكن تطوير ال business

الداتا التي لا توجد آلية لاسترجاعها فإن تخزينها أمر مغلوطة

المرحلة الثانية من دورة حياة البيانات هي التخزين (storage):

✓ نهتم في هذه المرحلة أن الداتا موجودة دائماً ونهتم فيها بال security لحد معين.

- ✓ نقوم بتخزين الداتا حتى تعمل عليها *process* من نمط معين.
- ✓ حيث يتم تخزين هذه البيانات عادة في مجموعة البيانات بتنسيق رقمي.
- ✓ يجب أن يكون تخزين البيانات غير متقلب ويجب أن يوفر قدرات على صيانة البيانات والوصول إليها واسترجاعها.
- ✓ يجب تنفيذ الإجراءات الأمنية من أجل فرض تناسق البيانات وحماية هذه البيانات.
- ✓ التركيز على استعادة البيانات والبيانات غير المهيكلة.

المرحلة الثالثة من دورة حياة البيانات هي المعالجة (*process*):

- ✓ في مرحلة ال *processing* نهتم جيدا بال *security*.
- ✓ ويتم فيها معالجة البيانات.
- ✓ ويتم فيها تنفيذ إجراءات الأمان لتجنب الوصول غير المصرح به إلى البيانات.
- ✓ ال *OLTP* مقابل *OLAP* كما ذكرنا الفرق بينهم سابقا.
- فهناك هرمية يجب مراعاتها دائما عند العمل مع ال *data* أو على الأقل كل من يعمل مع ال *data* يجب أن يكون مراعي لهذه المصطلحات:

➤ *Data source*:

- وهي مصدر هذه ال *data* حيث تكون موجودة من الطبيعة ولكن لم نحص عليها بعد .
- وعند الحصول عليها يصبح اسمها *data source* وحيث كما تحدثنا المصادر ممكن أن تكون *primary* أو *secondary* ممكن من الناس أو من الاستبيانات أو ... (من مصدرها مباشر أو من مصدر غير مصدرها).
- حيث تدخل البيانات إلى ال *database* إذا كنا نتحدث عن *structured data* , وإذا لم تكن *structured* فتذهب إلى *data set* , المهم تخزينها بمكان ما.
- ثم يصبح لدينا *pre processing* و *data integration* و *warehousing* المهم أن تصل لمرحلة أستطيع فيها أن استعملها .
- وبإبقاء *database* حية هي مهمة ال *DBA* (Data base administrator) ولا يهم طريقة استعمال هذه الداتا .

➤ (Data exploration):

على فرض أنه قد وصلتنا ال *data set* جديدة ونظيفة .

✓ مثال: علامات بكالوريا بدون أسماء وأردنا إخراج مانستطيع إخراجها من المعلومات فماتقوم به هو *exploration* وهو أن ننظر إلى هذه الداتا لنرى ماهي كأن ننظر إلى توزيع العلامات مثلا علامة مادة الرياضيات من 60 فنرى كم طلاب بين ال 0 وال 5 وبين 5 وال 10 وهكذا , وهل هذا التوزيع طبيعي أو يثير الاهتمام , ونعالج الداتا هل هناك داتا ناقصة أو خارجة عن الطبيعة مثلا العلامة 70 خارجة (الحد الأعلى هو 60).

➤ (data mining or information technology):

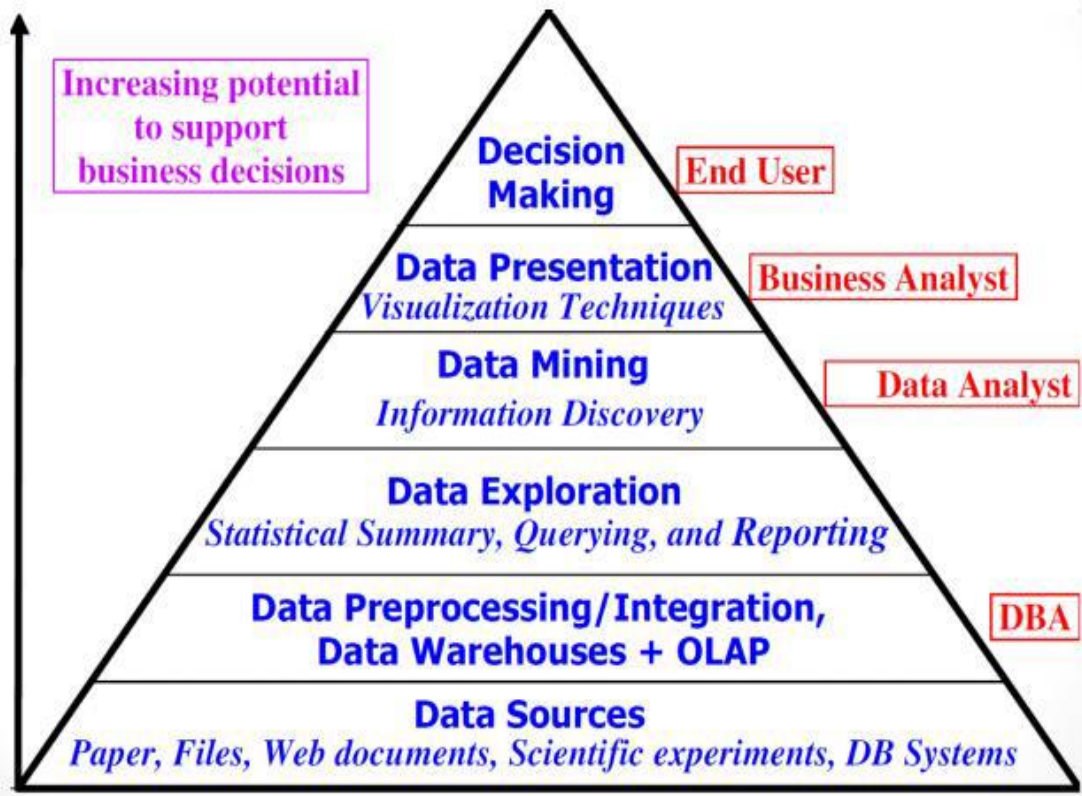
بعد التأكد من صحة الداتا وهي واقعية ومنطقية وقمت بـ *exploration* عليها ورأيت بعض العلاقات وتأكدت من نظافتها , نبدأ عندها بمرحلة ال *data mining* وهي مرحلة استخراج معرفة من هذه الداتا ونرى إذا فيها ترابط , أو نستطيع استخراج نتيجة مفيدة منها حيث نقوم بعمل *data analysis* للداتا الموجودة ولانضيف شيء من عنا بل نعتمد على ماتقولها الداتا فقط.

➤ (data presentation):

كيف نعرض هذه الداتا فليس من المعقول أن نعرض علامات 1000 طالب, بل يجب أن نجد طريقة تعبر عن المحتوى مثلا نرسم مخطط يوضح هذه الداتا, وعندها نعرضها *business analyst* وهو الشخص الذي سينظر إلى هذه المخططات ويددلي برأيه المتخصص مثلا أن يحدد سعر سوق العقارات أو أن يقول أنه يجب أن نركز على هذا المجال وما إلى ذلك.

➤ (Desion Making):

وهنا يأتي دور الشخص الذي سيأخذ القرار النهائي وعلى أي أساس نحن نتنقل شاقوليا في هذا الهرم؟
وهو على أنه اقتربنا من اتخاذ القرار , حيث أنه كلما صعدنا للأعلى كلما اقتربنا من اتخاذ القرار النهائي الأنسب .



المرحلة الرابعة من دورة حياة البيانات هي التوزيع والنشر (*dissemination*) للبيانات:

✓ وهي مرحلة توزيع الداتا حيث لا يتم إتلاف البيانات بشكل عام.

✓ نادرا ما يتم تجاهل البيانات بناء على طلبات معينة.

لا نتحدث عن تدمير الداتا إلا في حالات نادرة، فأحيانا في القانون يكون مطلوب منا أن ندمرها ولكن لانذكر هذا كجزء من حياة الداتا لأن يجب أن تكون موجودة دائما في مكان ما حتى لو لم نعد بحاجة إليها نقوم بنقلها أو بيعها لأحد آخر كي لا نضيع الجهد الذي بذلناه في جمعها وتنظيفها و... في هذه الحالة ماذا سنفعل بها إذا؟؟ يمكننا نقلها لجزء آخر من نفس الشركة لدعم القرار مثلا أخرجنا ال *data set* نظيفة بعد عمل *transaction* لمدة سنة عليها ونقلناها لقسم آخر لاستخراج قرارات منها أو أي شيء آخر، أو ننقلها لأشخاص يقومون بـ *knowledge creation* حيث يأخذون عدة *dataset* ويجمعونها سوياً أي يخرجوا داتا إحصائية عن السوق بأكمله.

Data-Quality(DQ):

✓ إن جودة البيانات ضرورية للتطبيقات كثيفة البيانات.

✓ أي *application* يبني على *data* لا يوجد فيها *quality* النتائج التي ستنتج ليس لها فائدة ومعنى.

✓ حيث أداء التطبيقات وصلاحيّة المعرفة المكتشفة تعتمد على صحة البيانات.

- ✓ عندما تصبح الأنظمة أكثر تعقيدا تصبح بيانات الجودة *DQ* محورية أكثر (والذي يحدث حاليا ويتم العمل فيه هو
- ✓ أن ال *data base system* أصبحت معقدة حتى التي فيها *structured data* , مثلا بأن يتم توزيع ال *Data* أي أن تصبح *parallel* وبالتالي أصبحنا نحتاج عدة آلاف معالجات وعدة آلاف *hard disk* , فإذا أردنا الحفاظ على جودة البيانات وهي على هذه الدرجة من التوزيع فمفهوم ال *data quality* نفسه سيصبح معقد.
- ✓ ينبغي مراعاة ال *DQ* في جميع مراحل دورة حياة البيانات (جمعها ومعالجتها وتوزيعها) , ليس فقط أثناء المعالجة المسبقة لجميع التطبيقات حيث تعرف ال *DQ* أثناء المعالجة المسبقة بتنظيف البيانات .
- ✓ أكثر ما نركز عليه هو ال *data cleaning* والسبب هو أن أغلب الناس يقومون بتجميع الداتا من أي مكان ومن كل المصادر المتاحة ثم يحللون فيما إذا كانت جيدة أم لا , أي أنه لا يقوم بتصميم أدوات تجمع داتا نظيفة بل تجمع داتا فقط , وبعد أن قام بتجميع هذه الداتا التي يتوقع أن يجد فيها المعلومة التي يريد بدأ ببناء ال *data set* وعندها يبدأ بتنظيف الداتا , فأغلب الوقت يصرف في التنظيف , حيث يتضمن تنظيف الداتا ملء القيم المفقودة , وإزالة القيم الخاطئة , وتمهيد البيانات الصاخبة و...., حيث ينفق مهندس البيانات ما يقرب من 90% من وقت تنظيف البيانات , ففي هذا الجزء لدينا مشكلة ويجب تحسينها بأننا عندما نجمع الداتا نحرص على أن تكون نظيفة , حيث بدون بيانات نظيفة يصبح التطبيق المكثف للبيانات عديم الفائدة , " يجب اتخاذ قرارات عالية الجودة باستخدام بيانات عالية الجودة " .
- ✓ ماهو المقصود بال *dirty data* ؟ هو أن يكون لدينا حقول فارغة أو حقول قيمها لا تطابق الواقع أو هناك *record* يوجد فيها *noise* أي يوجد ضجيج على مستوى الحقل بأكمله مثلا لدينا حقل من الحقول هو الدخل لمجموعة أشخاص وهذا لحقل قمنا بحسابه نحن دون سؤال الأشخاص المعنيين عندها هذا الحقل فيه ضجيج , أو أن ال *data insistance* أي أنه فيها تكرار وفي كل مرة يوجد قيم مختلفة مثلا بيانات شخص ما مكررة وفي كل مرة لديه قيمة عمر مختلفة , أو أن ال *data incomplet* أي أعلم أن الداتا لديها بقية ولكنها ليست موجودة لدي , أو أنها *out of date* مثلا شيء كان سعره هكذا من 10 سنين وهذه المعلومة ليس لديها قيمة الآن .
- ✓ حيث أن هذه الأخطاء متى تحدث؟؟ أثناء إدخال الداتا أو النقل وأحيانا لدينا عدة أجزاء من ال *application* وال *transaction* حصل في جزء ما دون الآخر .
- ✓ تحولت دراسة ال *data quality* من أن نقول أن هذه هي ال *measurements* التي يجب أن نقيسها إلى أن نقول أن هناك *guidelines* , أي ماهي النصائح التي يجب اتباعها , وهذا لا يعني أنه لا يوجد *measurements* والأهم أن نقول ماهي الإجراءات التي سنتبعها للحفاظ على نظافة الداتا.
- ✓ تضمين عمليات التحقق من البيانات خلال مراحل مختلفة من النظام.

✓ إجراء فحوصات اتساق البيانات عند دمج البيانات الجديدة.

✓ إجراء فحوصات سلامة بشكل روتيني.

✓ الإبلاغ عن البيانات المشبوهة.

✓ ال *guidelines* : هي القيام بعملية *check* في كل مرحلة , ويمكن أن يكون أوتوماتيكيا , مثلا هذا الحقل يجب أن يكون بين [0,10] وأي قيمة خارج المجال نطلب من المصدر إعادة إرساله .

✓ Data consistency check : يجب أن نتأكد أن الداتا متوافقة مع بعضها , مثلا لدينا معلومات عن شخص معين أن راتبه 20 ثم حصلنا على معلومة عن الشخص نفسه أن راتبه أصبح 200 فنلاحظ انزياح فهناك خطأ في المعلومة ويجب مراجعتها , وخاصة بالداتا الموزعة يجب أن تكون متوافقة .

✓ Sanity check : نتأكد من منطقية الداتا وعند حدوث أي خطأ في أي مرحلة يجب أن نقوم بعمل *flag* له (أي أما أن نحذفه أو نقوم بتعبئته أو).

❖ Measurements : لدينا داتا ونريد قياس جودتها :

✓ Accuracy : تعني دقة البيانات أي إذا كان الواقع الخارجي الذي أحضرت منه هذه القيمة لديه نفس القيمة التي لدي هذا يعني أن الداتا لدي دقيقة ولدينا نوعان من ال Accuracy:

✚ Syntactic : وهو أن نمط الداتا منطقي أو لا .

مثال: شخص معين اسمه *john* فإذا كان مكتوب كما هو فهذا منطقي أما اسمه كرقم فهذا غير منطقي ويظهر *Syntactic error* وهي سهلة الاكتشاف ويمكن هذا ال *check* إضافتها على ال *database or xml*.

✚ Semantics : أي هذه الداتا تطابق الواقع أم لا أي موجودة لدي .

✓ Completeness : وهي أن المعلومات لدي كاملة .

✓ Consistency : وهي إيجاد تعارضات في الداتا .

✓ Timeliness : وهو الوقت الذي نحتاجه لتغير سعر قطعة ما مثلا في الداتا .

✓ *value added* : هذه المعلومات هل لديها قيمة.

✓ Interpretability : هل يستطيع الإنسان أن يفسرها .

✓ access ability : وهو يعني اعتقاد البعض إنه ليس بالضرورة البيانات التي لها قيمة هي التي يمكننا الحصول عليها وقد يكون هناك بيانات لا يمكن الحصول عليها لكن لها قيمة , أي يكفي معرفتنا بوجود هذه البيانات ليكون لها قيمة.

أنواع ال application:

- **olap (Data Analytics Applications)**: كما في *database* وأمثلتها هي أن نقوم ب *summaries* ونرى الناس إذا كان لديهم *satisfaction* أو إذا كان لدينا *abnormality*, وغالبا مايستخدم نماذج نجمة متعددة الأبعاد (مكعبات البيانات) .
 - **KDD (Knowledge Discovery Applications)**: لنفرض أنه لدينا *dataset* وأنه بطريقة ما استطعنا أن نوصلها إلى *structure* يمكننا إخراج قيم منها , ونسميها *pattern analysis* ونرى القيم التي تتكرر , وتشمل معظم المهام الشائعة :
 - **Association;**
 - **Classification;**
 - **Clustering;**
 - **Anomaly detection; and**
 - **Recommendation.**
 - **DSS (Data – intensive applications)**: وهي عبارة عن *database* كبيرة ولها واجهة فقط , كما *student information system* وهناك *meta design* مشترك بين هذه الأنواع.
 - **Data Flow (Decision making systems)**: ويعني أنه بعد تحليل ال *better* يجب أن نستخرج قرار منها فنحتاج إلى *dashboard* أو *chart* ليعطي نتيجة , وهو الشخص الذي ليس له علاقة بال Technology أبدا.
- ملاحظة: يمكن الاستفادة من مراجع الدكتور الموجودة في نهاية الساید الأول .

Types of Data Analysis

لدينا فرضا داتا مخزنة ضمن نظام ما وعلينا استنتاج معلومات منها، هذا الاستنتاج يتم بأحد الأنواع التالية:

- **Descriptive:**
أي توصيف الداتا، توصيف كمية الميزات الرئيسية لها، ويمكن من خلال التوصيف الوصول الى بعض المعلومات، العملية الأساسية هنا هي **description**
- **Exploratory:**
وتعني تحليل الداتا من اجل اكتشاف معلومات لإيجاد روابط مسبقه غير معرفة فيها (علاقات عامة) ويمكن فعل ذلك باستخدام أدوات تحليل وتشكيل أسئلة تخص الداتا
- **Inferential:**
وهي علاقة dataset الى dataset، أي هذه العينة من الداتا الى أي مدى تتناسب مع العينة الأخرى وما مدى الاختلاف بينها وإيجاد علاقات إحصائية بينها لمعرفة كمية الارتباط أي اذا تغيرت قيمة احدها هل ستؤدي الى تغير الأخرى ؟
- **Predictive:**
التعلم عن طريق داتا معينة من اجل القيام بعمليات تنبؤ بناء على المعرفة الناتجة عنها، التنبؤ ليس بالضرورة للمستقبل ويمكن استعماله لتوقع نتيجة بتغيير احد المعطيات في الماضي (لو حصل x كان سيحصل y)
- **Causal:**
إيجاد السبب الذي يجعل الداتا تتغير أي ان هذه المعلومة ستؤدي الى هذه المعلومة

Descriptive Data Analysis

هي عملية بديهية مثل تحديد (Definition of each table, column , Database ERD) وقياس quality of data , والبحث عن الداتا المفقودة , او الداتا المنتهية الصلاحية او ذات القيم الفارغة
Rudimentary charts and visualization

مثال:

بفرض لدينا هذه الداتا

	A	B	C	D	E	F	H	I	J	K
1	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	AVG_DURATION	SATISFACTION	USAGE_LEVEL	CONSIDERING_CHANGE_OF_PLAN
2	zero	31953	0	6	313378	161		4 unsat	little	no
3	one	36147	0	13	800586	244		6 unsat	little	considering
4	one	27273	230	0	305049	201		15 unsat	very_little	perhaps
5	zero	120070	38	33	788235	780		2 unsat	very_high	considering
6	one	29215	208	85	224784	241		1 very_unsat	little	never_thought
7	zero	133728	64	48	632969	626		2 unsat	high	no
8	zero	42052	224	0	697949	191		5 very_unsat	little	actively_looking_into_it
9	one	84744	0	20	688098	357		5 very_unsat	little	considering
10	zero	38171	0	7	274218	190		5 very_sat	little	actively_looking_into_it
11	zero	105824	174	18	153560	687		4 very_sat	little	never_thought
12	zero	20120	43	0	623166	209		8 very_sat	little	never_thought
13	one	50939	76	13	587207	336		5 avg	little	considering
14	zero	23553	244	0	926178	158		5 very_unsat	very_little	actively_looking_into_it
15	one	143501	63	0	515444	530		10 unsat	high	considering
16	one	36940	259	0	979303	236		8 very_sat	very_high	considering
17	zero	159902	0	20	213299	516		4 very_sat	high	actively_looking_into_it
18	zero	45482	82	6	607518	157		2 avg	very_high	no
19	zero	41513	0	15	214276	167		2 very_unsat	very_high	considering
20	zero	53391	74	69	599957	287		2 very_unsat	very_high	no
21	one	52308	0	50	187864	394		2 very_unsat	little	considering
22	zero	129795	0	0	281839	684		13 very_unsat	very_little	no
23	one	86658	78	15	150852	296		5 very_sat	very_high	never_thought

من البديهيات النظر الى كل سطر على انه حالة تخص مكون واحد ف بافتراض ان الداتا تعبر عن حالات بيع اذاً بعدد عدد الاسطر يمكننا معرفة عدد حالات البيع
 وإذا كان كل عمود يعبر عن معلومة ف بعدد عدد الاعمدة ينتج عدد المعلومات المسجلة عن كل حالة
 اصطلاحاً:

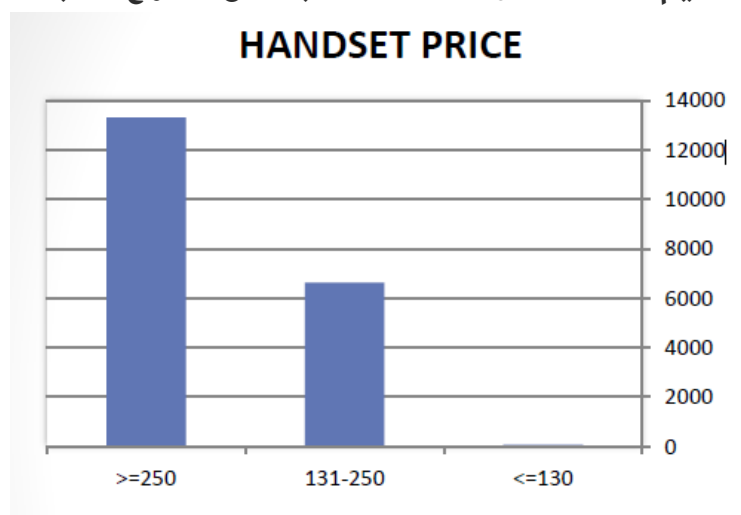
✓ كل سطر يدعى example او sample او object

✓ كل عمود يدعى feature او attribute او قياس او data measurity

يمكن عمل التوصيف بشكل صور visualization ,

مثال 1:

بتقسيم أحد أعمدة ال dataset السابقة الى 3 أنواع حسب price

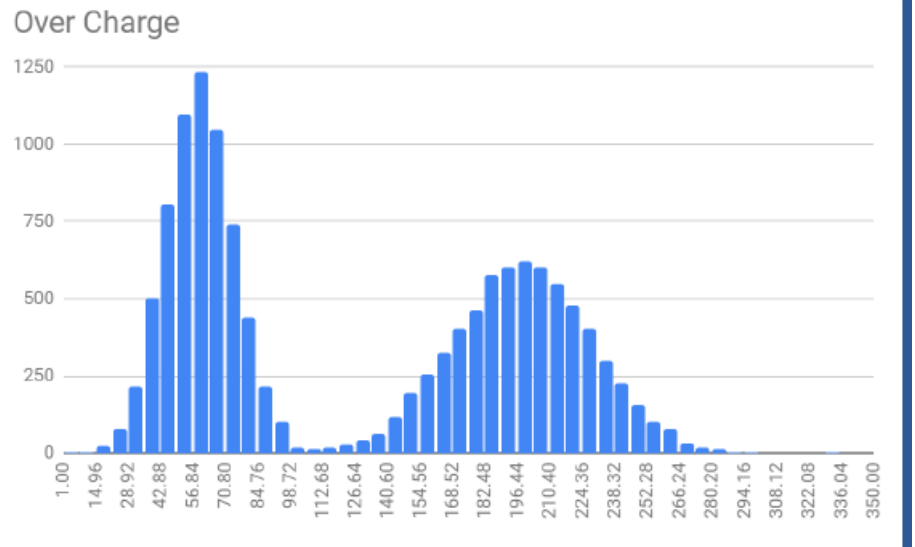


باعتبار عدم المعرفة المسبقة للمسألة بالإمكان الاستنتاج من الشكل كبدائية ان العينة تحوي 14 ألف وهو كرقم يعبر عن المجتمع، اذا يمكن الاستنتاج أيضا ان المعلومات إحصائية ومن هذه المعلومات فإن ثلثها مصنف ضمن category والثلث في category والقليل جدا في اخر category فإذا كانت من رغبات الإدارة بما يخص المادة التي تم رسم البيان لها هي الزيادة فمنطقيا يجب توجيه الزيادة في القسم الذي سيحوي أكبر عينة مستفيدة .

- الداتا السابقة هي داتا استهلاك لـ واحدات الهاتف وقد تم اخذ عدد من القياسات من اجل الحصول على معلومات منها
 - الرسم البياني يمثل اعداد الأشخاص التي تحمل أجهزة عالية، فضمن المجتمع الثلثين يحمل أجهزة عالية وثلث متوسط والقليل دون المتوسط وبالتالي هكذا معلومة قد تفيد الباعة للتركيز على استهداف فئة معينة
- ملاحظة: قرار التوزيع في category واختيار كل category ليس ثابتا , أي ان كل شخص قد يقوم باختيار ثوابت أخرى للمقارنة مثلا

مثال 2:

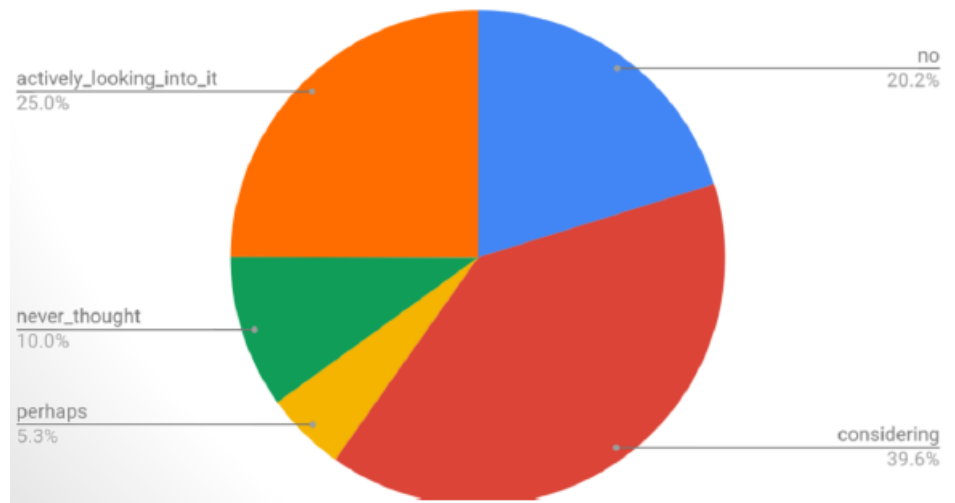
في الصورة description لعمود اخر من dataset السابقة



تظهر الإحصائية نسبة الأشخاص الذين يتجاوزون الحزمة الخاصة بهم وحسب chart السابق يمكن استنتاج وجود نوعين للأشخاص، اما اشخاص بحزمة كبيرة ولا يقومون بتجاوزها او ان الحزمة خاصتهم صغيرة

تذكير: ما زلنا نقوم بتوصيف الداتا ولكن خلال توصيف الداتا يمكن استنتاج معلومة.

Considering Change of Plan



في الشكل pie chart:

لدينا خمس أنواع كما هو واضح بعض الأنواع أكثر من غيرها، إذا كان ال chart يعبر عن صحة الشركة فمثلاً يجب القيام باستهداف العناصر التي تزيد من صحتها

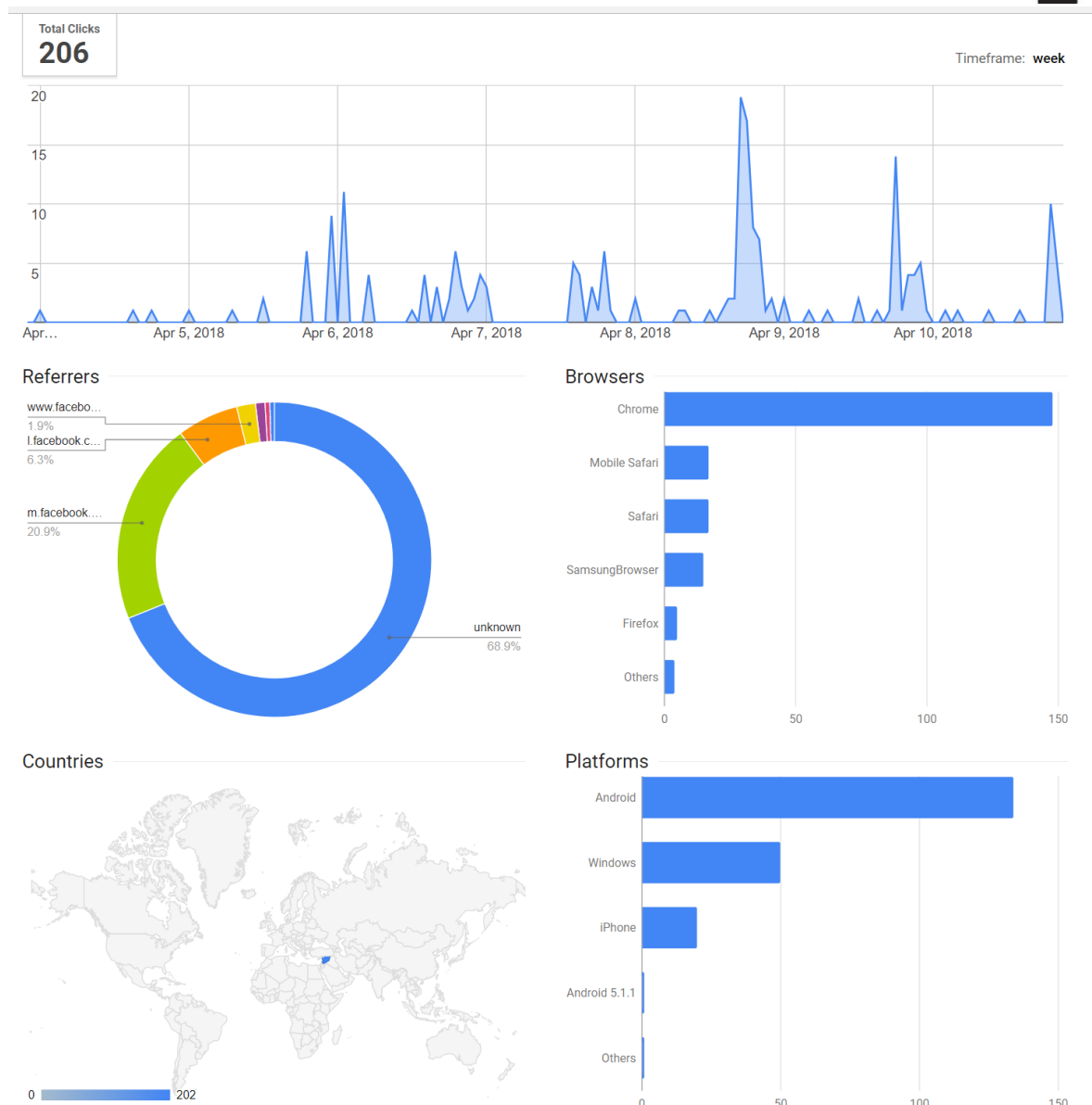
الشكل يعبر عن آراء الموظفين إذا قاموا بالتفكير في ترك الشركة وتم التصنيف في خمس أنواع: no , considering never thought , perhaps , actively looking into it

اللون الأحمر يعبر عن الأشخاص الذين لم يعط إجابة واضحة لذلك من صحة الشركة استهدافهم من أجل تحصيل ولائهم لها:

➤ بناء على أننا نقوم بالكثير من description للبيانات تم الوصول إلى مفهوم dashboard وهي مجموعة مخططات live تعبر عن المخططات خلال الوقت الحالي

في الصورة مثال عن dashboard

توصف مجموعة الأشخاص الذين يدخلون إلى موقع ما، ما اسم المتصفح، البلد الذي تم تسجيل الدخول منه ...



Exploratory Data Analysis

Exploratory: تبحث في العلاقات بين الداتا، أي محاولة اكتشاف علاقة بين نوعي داتا

- Analysis of trends, changes, statistical anomalies, and previously unknown relationships
- Looking at distributions of feature values
- Looking at correlations between features
- Feature wide analysis
- Generally looking for something out of the ordinary or interesting

لدينا مجموعة علامات لمادة، الامتحان الخاص بها كان يحوي 4 أسئلة، ف dataset تحوي علامة كل سؤال والمجموع الخاص بها:

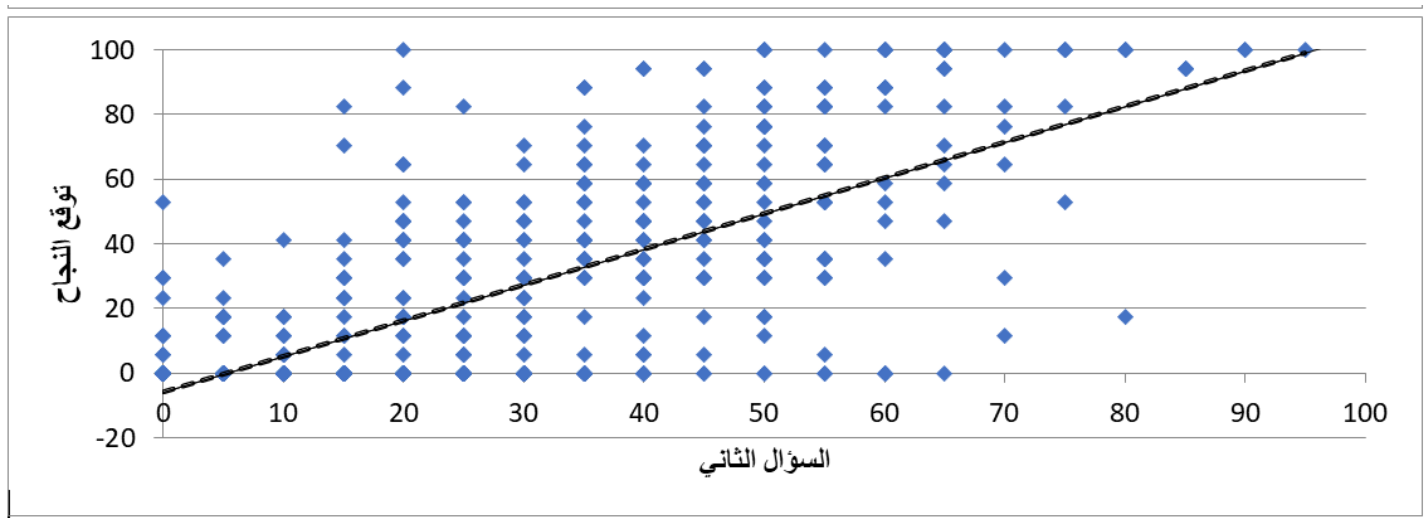
بافتراض كان عدد المتقدمين 300 فيكون عدد التسجيلات 300 ، كانت علامة السؤال الأول 10 ، الثاني 20 ، الثالث 10 ، الرابع 30 ويكون المجموع 70 ، ومن الداتا المبينة ينتج الجدول:

نلاحظ ان هذه الداتا إحصائية وما تم القيام به هنا هو توصيفها

11	عدد الأصفار
64	أعلى علامة
560	عدد الكل
1	عدد فوق 60
6	عدد فوق 55
16	عدد فوق 50
41	عدد فوق 45
88	عدد فوق 40
5.446429	وسطي سؤال 1
4.398214	وسطي سؤال 2
4.419643	وسطي سؤال 3
11.71786	وسطي سؤال 4
25.98214	وسطي مجموع
13.23016	انحراف معياري

E	D	C	B	A	
علامة	Q4	Q3	Q2	Q1	1
44	16	8	14	6	2
39	17	10	6	6	3
33	11	10	4	8	4
21	7	8	0	6	5
10	6	0	0	4	6
36	15	8	5	8	7
24	14	0	2	8	8
25	12	0	3	10	9
22	12	0	2	8	10
22	16	0	0	6	11
28	0	10	12	6	12
36	15	8	9	4	13
40	20	8	8	4	14
29	17	0	2	10	15
27	13	8	0	6	16
34	13	5	8	8	17
23	9	6	2	6	18
27	9	9	5	4	19
11	1	2	0	8	20
8	2	0	0	6	21
27	15	8	0	4	22
37	19	7	9	2	23
21	6	7	2	6	24
19	7	6	0	6	25
21	14	2	1	4	26

إذا قمنا بأخذ علامة السؤال الثاني وتحويلها الى مقياس من 0 الى 100 , وكذلك علامة النجاح , بتحديد النقاط على الرسم ينتج لدينا (علامة السؤال الثاني على محور X ومجموع العلامة على محور Y)



(مقياس علامة النجاح من 0 الى 100 يعبر عن توقع النجاح)

بالنظر الى توزع النقاط يمكن ملاحظة علاقة بين توقع النجاح وبين العلامة في السؤال الثاني (الطالب الذي علامته في السؤال الثاني عالية اكثر من 50% نلاحظ ان توقع النجاح له اكبر) العلاقة ليست واضحة بشكل كبير ولكن بالقياس (احد متغيرات القياس يدعى correlation عند حسابه ينتج $correlation > 60\%$) أي ان السؤال الثاني يؤثر بشكل كبير في احتمالية النجاح .

مثال 2

في عينة ثانية:

Rules

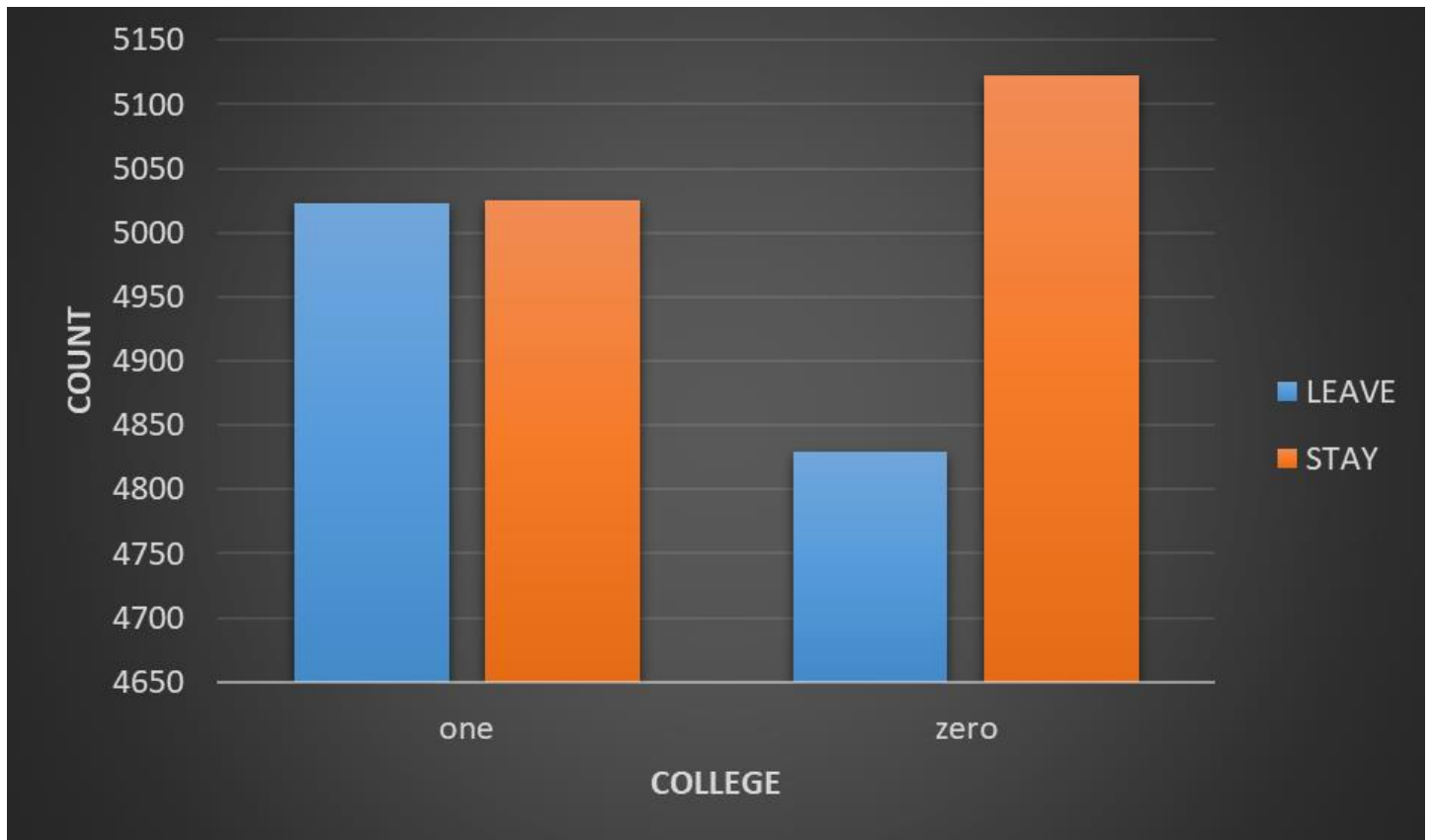
Rule	Support	Confidence
Outwear \Rightarrow Hiking Boots	33%	66.6%
Outwear \Rightarrow Footwear	33%	66.6%
Hiking Boots \Rightarrow Outwear	33%	100%
Hiking Boots \Rightarrow Clothes	33%	100%

تقرأ بالشكل من قام بشراء الغرض في الطرف اليسار قام بشراء غرض الطرف اليمين وهذا حصل بنسبة 33% انه تم شراء الغرضين معا

ونقوم بحساب العلاقة بين هاتين ال features (Hiking Boots, outdoor) وهذه أيضا عملية exploration على الرغم ان 33% هي نسبة صغيرة لكنها بالنهاية نسبة وتعني وجود ترابط حتى لو لم يكن قويا .

مثال 3:

بالنسبة للمحور x هناك توزيعين اما صفر او واحد



اذا كان واحد فالأحمر والازرق يتساويان، اذا كان صفر يختلف التوزيع، وهذا يعني ان المحور x يؤثر بالتوزيع باختلاف قيمته .

وهناك أيضا من عمليات exploratory عملية clustering سيتم الحديث عنها بالتفصيل لاحقا، وهي ان يكون لدينا في الفضاء ثم تقسيمها الى أنماط بناء على عدة عوامل (قرب ، شبه) ، فهي أيضا عملية استكشاف للداتا اذا طلب استكشاف الداتا فان العمليتين اللتين نقوم بهما هما التوصيف والاستكشاف .

Inferential Data Analysis

مقارنة عينتين، وهي من اجل معرفة هل العينات من ذات المجتمع ام هي مستقلة

وتعني انه مثلا عند وجود عيتين هل ستقومان بالتصرف بذات الطريقة تجاه منتج معين ام سيكون هناك اختلاف , التشابه في التصرف سيعني ان العينات من ذات المجتمع

لنفرض قمنا بطرح 5 منتجات في دمشق وحلب وتم القياس انهما من نفس المجتمع بالنسبة لهذه المنتجات بذلك غالبا طرح المنتج السادس سيتم التعامل معه بنفس الطريقة بالنسبة للعيتين , اما اذا كان هناك اختلاف في الازواق بالنسبة ل 5 منتجات فغالبا طرح المنتج الجديد سينتج عنه اختلاف

أي ان inferential يعمل على اكثر من dataset

Analysis of hypothesis

What changes if

Looks at statistics before and after test

Feature change analysis

Often aims to measure similarity of sample to real world.

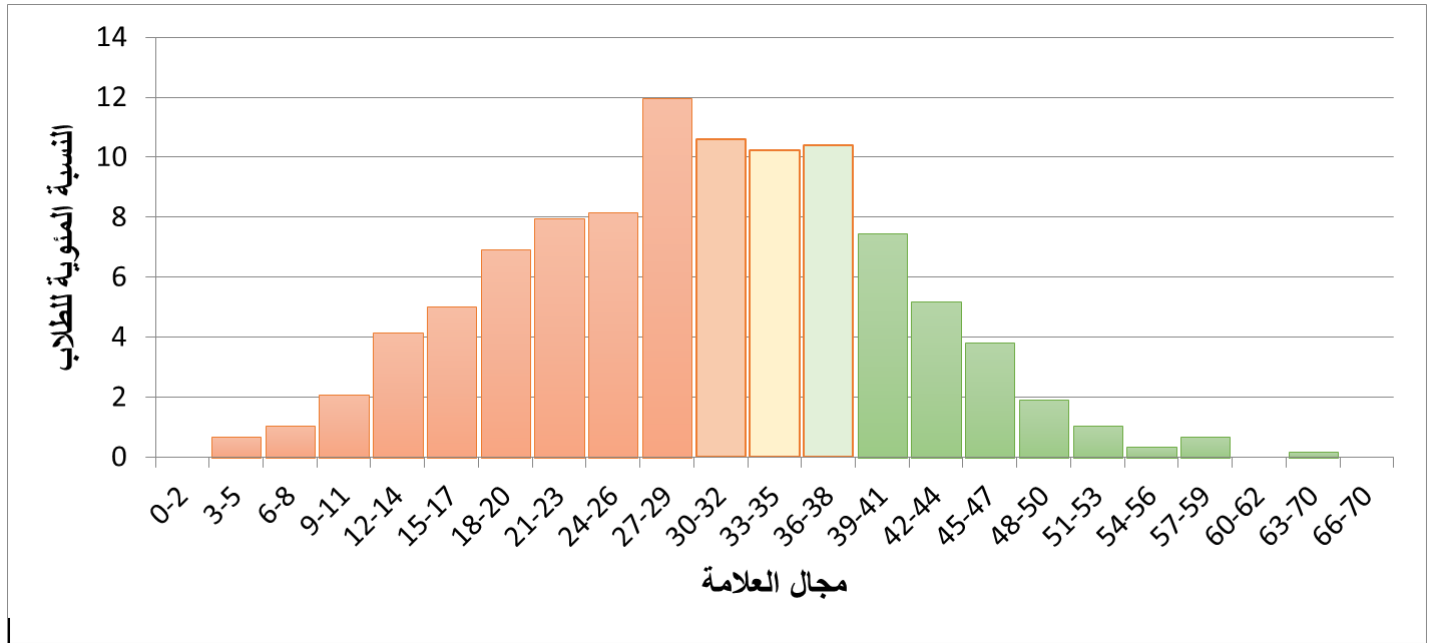
Generally looking to prove/disprove a hypothesis

مثال :

بالعودة الى dataset الخاصة بعلامات الطلاب



الصورة السابقة قبل القيام بالتدخل حيث انهم نفس الطلاب وتم اجراء امتحان لهم مع اختلاف في الأسئلة



بالنظر الى النتيجة بعد التدخل نلاحظ ان الطلاب ليسوا مجتمع واحد لأنه اثر في مجموعة منهم ولم يؤثر بالأخرى (التدخل كان بإزالة الاختيار من متعدد من الأسئلة 3 :)

المقاييس التي نستطيع استعمالها :

من المهم معرفة ان عملية ما مثلاً قابلة للقياس عن طريق احد هذه المقاييس

فعن طريق القياس استطعنا ان نقول مثلاً انه بنسبة 70٪ لدينا عينتين

وهو بالنهاية يشكل مفهوم ال inference عن طريق القيام بتنبؤ احصائي على العينات التي لدينا

من المقاييس correlation : ابسط مقياس لعينتين احاديات البعد (نسبة بينهما)

Common Statistical Tests

Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means: <i>look for the difference between the means of variables</i>	
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression: <i>assess if change in one variable predicts change in another variable</i>	
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric: <i>are used when the data does not meet assumptions required for parametric tests</i>	
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

Predictive Data Analysis


التنبؤ مرتبط في ذهننا بالمستقبل ولكن يمكن اعتماده بطرق أخرى , مثل لو اننا قمنا بعرض ما على هذا الشخص العام الماضي فما هي ردة فعله , اذا ان أي حالة لا يوجد الخرج الخاص بها ضمن الداتا يعتبر تنبؤ عادة يتم ارجاع مسائل predictive الى هذا الشكل :

$$f(\text{features}; \text{parameters}) \rightarrow \text{Label}$$

حيث label هي القيمة التي أحاول التنبؤ بها
 وإذا نحاول إيجاد تابع معين يربط القيمة الخاصة بمسألتنا بنتيجة ما (label) , عن طريق المعلومات الموجودة وتدعى هذه الطريقة supervised learning , فكلما زادت الداتا كلما صارت المسألة اسهل

مثال

من dataset العلامات السابقة اخذنا علامة السؤال الثاني مع علامة السؤال الرابع



	Pass	Q4	Q2	A
1	1	16	14	2
2	1	17	6	3
3	0	11	4	4
4	0	7	0	5
5	0	6	0	6
6	1	15	5	7
7	0	14	2	8
8	0	12	3	9
9	0	12	2	10
10	0	16	0	11
11	0	0	12	12
12	1	15	9	13
13	1	20	8	14
14	0	17	2	15
15	0	13	0	16
16	0	13	8	17
17	0	9	2	18
18	0	9	5	19
19	0	1	0	20
20	0	2	0	21
21	0	15	0	22
22	1	19	9	23
23	0	6	2	24
24	0	7	0	25

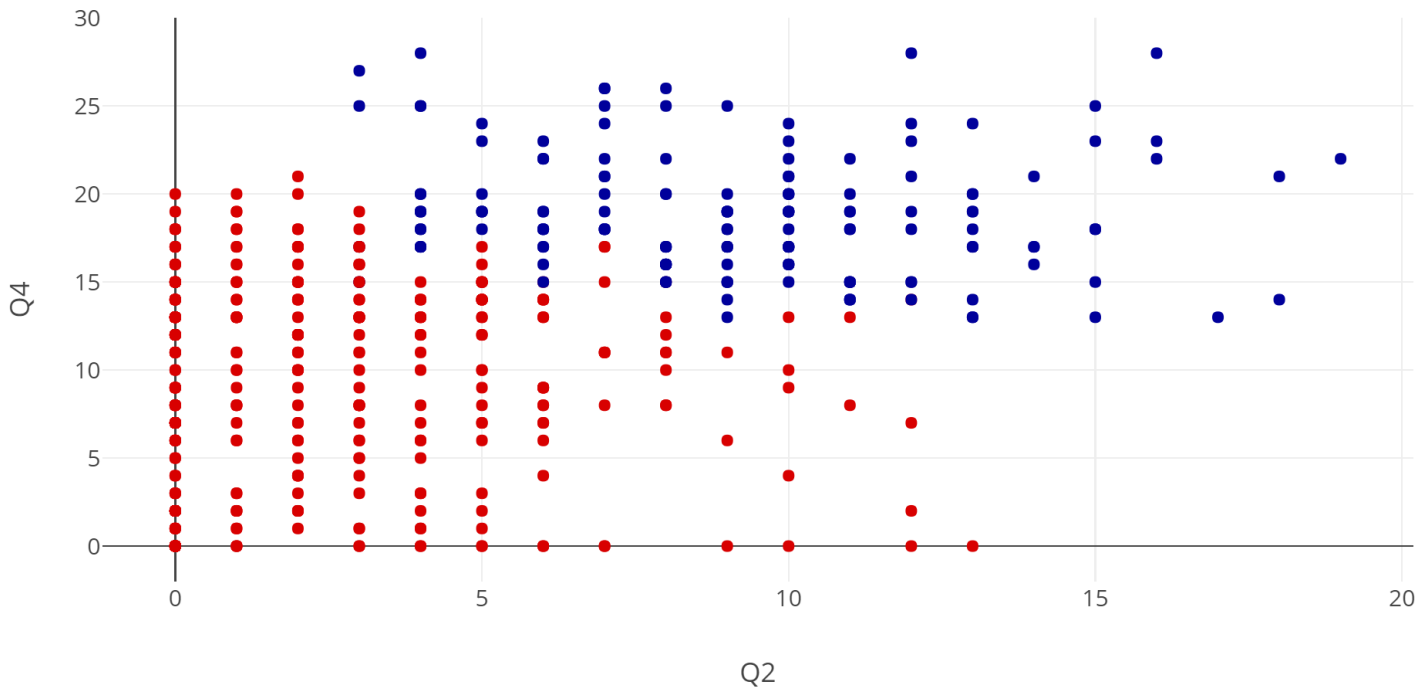
إذا تم طرح سؤال : هل يوجد ترابط بينها ؟ تكون الإجابة بالاعتماد على **exploration**

كم عدد العينات ؟ الإجابة ب **description**

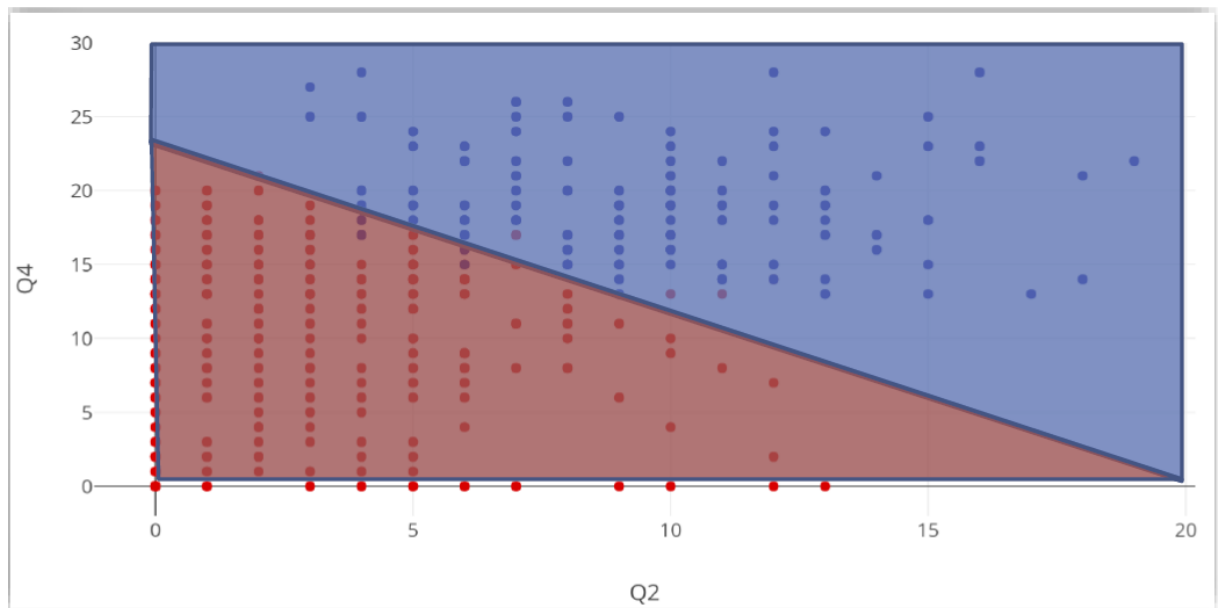
هل هناك فرق بين الطلاب الذين اجابوا عن السؤال الرابع والذين اجابوا على السؤال الثاني؟ , نفس العينة ؟ فالاجابة تحتاج **inference**

في **predictive** نحاول الحصول على معادلة تربط بين علامة السؤال الثاني والرابع واحتمالية النجاح فاذا وصلنا الى هكذا معادلة صار بالإمكان توقع النجاح او الرسوب بناء عليهما

نلاحظ أولا ان الطلاب الذين حصلوا على علامات عالية في السؤال الثاني كذلك حصلوا على علامة عالية في السؤال الرابع



يبيّن اللون الأزرق الطلاب الناجحين والاحمر الراسبين لذا وضوحا نتبين ان كلتا العلامتين تؤثران في النجاح فكلهما لهما نفس الوزن تقريبا , وما نحاول الوصول اليه هو الخط المميز في الشكل والذي يدعى **dissension** وهو يعبر عن معدلة رياضية خطية توضح الارتباط بين المعاملين حيث اذا كانت القيمة فوق حد معين فالطالب ناجح اما اذا كانت تحتها يكون راسب , تبقى المشكلة في العلامات التي تقارب الخط حيث لا يمكن الجزم بالنسبة لها



Label هو حالة خاصة من feature , اذا بالامكان عمل معادلة لهذه ال feature في المسألة التالية نحاول التنبؤ اذا ما كان الزبون سيبقى ام سيتترك المؤسسة وذلك بأخذ عينة عشوائية من الزبائن ومراقبة سلوكهم لمدة 6 اشهر وتحديد من ترك المؤسسة ومن لا و بايجاد علاقة مناسبة بين features بالامكان التنبؤ بسلوك الزبون لاحقا

	A	B	C	D	E	F	H	I	J	K	L
1	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	AVG_DURATION	SATISFACTION	USAGE_LEVEL	CONSIDERING_CHANGE_OF_PLAN	LEAVE
2	zero	31953	0	6	313378	161	4	unsat	little	no	STAY
3	one	36147	0	13	800586	244	6	unsat	little	considering	STAY
4	one	27273	230	0	305049	201	15	unsat	very_little	perhaps	STAY
5	zero	120070	38	33	788235	780	2	unsat	very_high	considering	LEAVE
6	one	29215	208	85	224784	241	1	very_unsat	little	never_thought	STAY
7	zero	133728	64	48	632969	626	2	unsat	high	no	STAY
8	zero	42052	224	0	697949	191	5	very_unsat	little	actively_looking_into_it	STAY
9	one	84744	0	20	688098	357	5	very_unsat	little	considering	STAY
10	zero	38171	0	7	274218	190	5	very_sat	little	actively_looking_into_it	STAY
11	zero	105824	174	18	153560	687	4	very_sat	little	never_thought	LEAVE
12	zero	20120	43	0	623166	209	8	very_sat	little	never_thought	STAY
13	one	50939	76	13	587207	336	5	avg	little	considering	STAY
14	zero	23553	244	0	926178	158	5	very_unsat	very_little	actively_looking_into_it	STAY
15	one	143501	63	0	515444	530	10	unsat	high	considering	STAY
16	one	36940	259	0	979303	236	8	very_sat	very_high	considering	STAY
17	zero	159902	0	20	213299	516	4	very_sat	high	actively_looking_into_it	STAY
18	zero	45482	82	6	607518	157	2	avg	very_high	no	STAY
19	zero	41513	0	15	214276	167	2	very_unsat	very_high	considering	STAY
20	zero	53391	74	69	599957	287	2	very_unsat	very_high	no	LEAVE
21	one	52308	0	50	187864	394	2	very_unsat	little	considering	STAY
22	zero	129795	0	0	281839	684	13	very_unsat	very_little	no	LEAVE
23	one	86658	78	15	150852	296	5	very_sat	very_high	never_thought	STAY

بالإمكان أيضا العودة الى الداتا المخزنة مسبقا والاستنتاج منها والتنبؤ خلال هذه الفترة، وهنا نتبين عمل data engineer فاذ كان لدينا quality data نستطيع استنتاج معادلة والتنبؤ بشكل صحيح اما إذا لم تكن كذلك فالمعلومات الناقصة ستؤثر في جودة التنبؤ ومعامل الثقة في المعادلة.

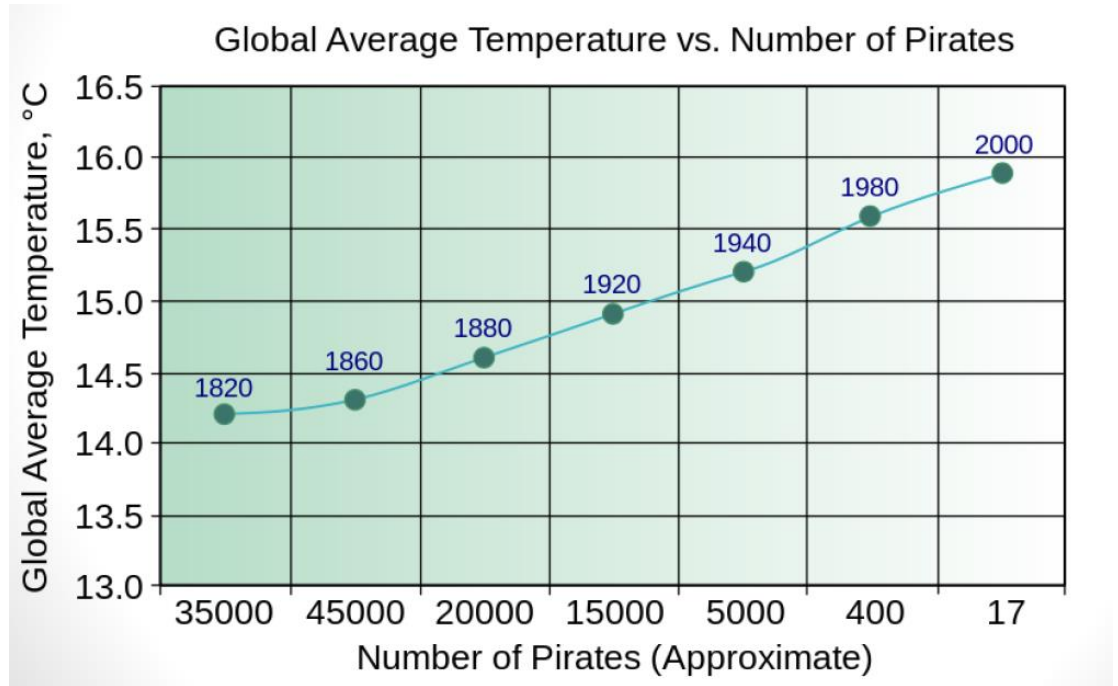
Predictive data analysis takes many names

- Classification: for single labels
- Auto tagging: for multiple labels
- Regression: for continuous targets
- Multi Regression: for multiple continuous targets
- Ranking Prediction: for ranked targets (i.e., IR tasks)
- Structure Prediction: for more complex structure

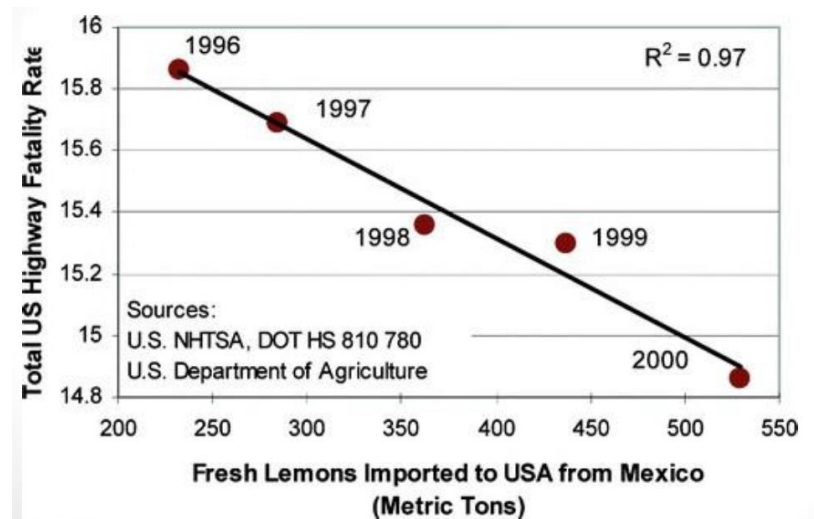
Data Analysis Causal

عند الوصول الى مرحلة هذه النتيجة بسبب هذا الامر، فتحليل الداتا هنا يكون قد اعطى value كبيرة الى صاحب القرار والوصول الى هذه المرحلة صعب جدا لأنه بحاجة الى تأكيدات

حيث يجب الاخذ بعين الاعتبار ان الارتباط لا يؤدي الى السببية **Correlation is not Causation** مثال 1



مع اعتبار عدم المعرفة ما هو x وما هو y نرى انه يوجد ارتباط بين المحورين فزيادة الأول يزداد الثاني لكن ال x تعبر عن عدد القراصنة في العالم ولا تعبر عن وسطي درجات الحرارة في العالم والنقط هي السنوات نلاحظ انه لا يمكن ان يؤثر بالعنصر الاخر هنا، أي ربما يكون هناك ارتباط ولكن السبب خارج هذه الداتا. مثال 2:



• المحور x يعبر عن كمية الليمون الذي تستورده الولايات المتحدة الأمريكية من المكسيك والمحور y يعبر عن عدد الحوادث القاتلة ويمكن الملاحظة انه يوجد ارتباط على البيان لكن لا يمكن إيجاد المسبب منها

- Causal analysis is interested in the root causes of observations
- For instance, what happens in a world not so similar to ours
Prediction in a counterfactual world.
- If we understand the causes in certain cases, we can answer such causal question.

Mechanistic Data Analysis

النمط السادس وهو يقيس درجة التأثير أي بعد إيجاد المسبب يقوم بقياس الى أي مدى يؤثر السبب في النتيجة وهو يعمل خارج الداتا تقريبا بالاعتماد على التحليل

Concerned with levels of change in underlying features

needed in order to cause a targeted change in the label.

Often incurs high levels of modelling and equation setting.

Resulting in parameterized models of the real world.

Includes Causal Analysis by default.

Example: How much screening would a population need in order to lower mortality rates by x%?

انتهت المحاضرة

