dm24s1

Topic 02 : Motivating Example

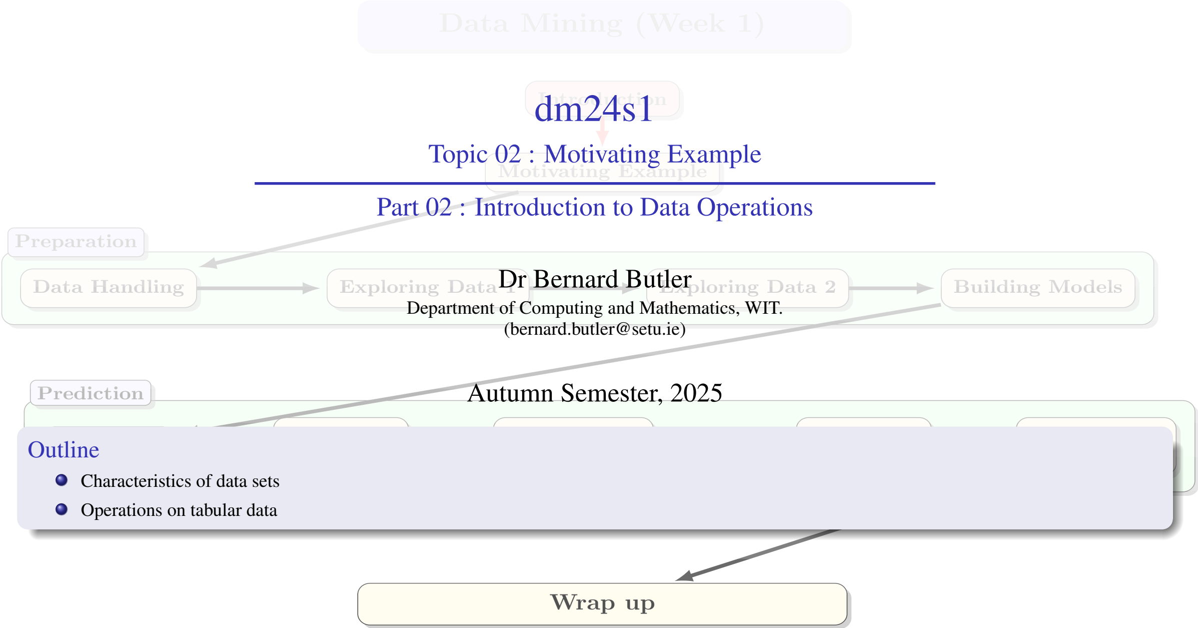
Part 02 : Introduction to Data Operations

Dr Bernard Butler

Department of Computing and Mathematics, WIT.

(bernard.butler@setu.ie)

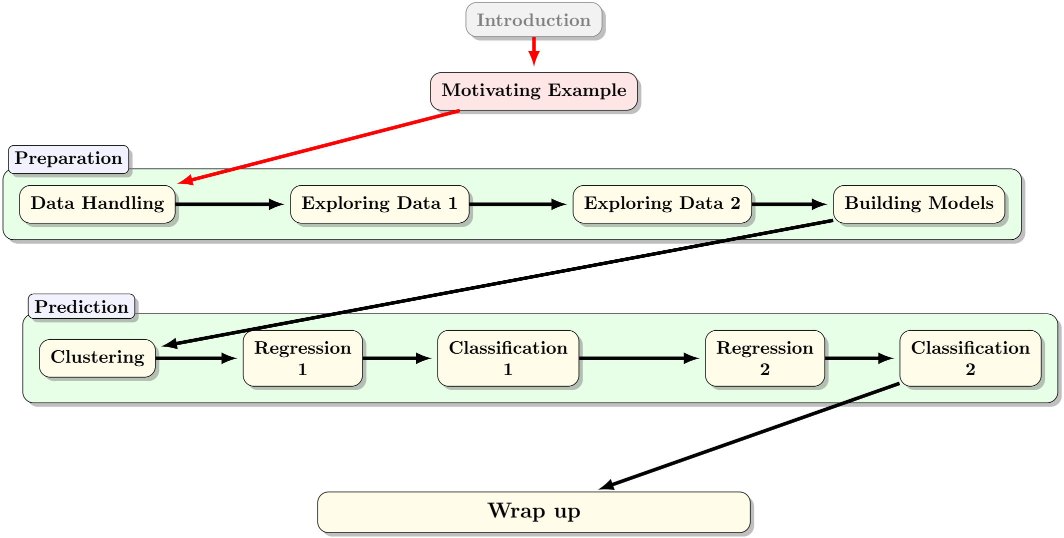
Autumn Semester, 2025



Outline

- Characteristics of data sets
- Operations on tabular data

Data Mining (Week 2)



Outline

Data sources

Type	Format	Example	DBMS	Language	Readiness for ML
Relational	Table	Transactions	MySQL, Post-gresql, ...	SQL	Maps to dataframe
Flat	Key + Value	Caches	Redis, mem-cached, ...	DBMS-Specific	Not rich enough
Document	Serialised objects	Tweets	Mongodb, Cassan-dra. ...	MQL, CQL	Too rich
Graph	Nodes and edges	Social relation-ships	Neo4j, Dgraph, ...	Gremlin, Cypher, DQL, ...	Specialised analyses
Columnar	DataSet	Logs	HBase, Spark DataSet	Hive QL, Spark SQL	Maps to dataframe

Preparing data

Data Preparation is the first step in data mining

In practice, data can be

- structured or unstructured,
- consolidated or scattered,
- consistent or inconsistent,
- clean or with error.

ML prefers structured, consolidated, consistent data, as clean as possible.

The `auto_mpg.csv` dataset already has these characteristics.

Outline

The auto-mpg dataset

1	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
2	18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
3	15	8	350	165	3693	11.5	70	1	buick skylark 320
4	18	8	318	150	3436	11	70	1	plymouth satellite
5	16	8	304	150	3433	12	70	1	amc rebel sst
6	17	8	302	140	3449	10.5	70	1	ford torino
7	15	8	429	198	4341	10	70	1	ford galaxie 500
8	14	8	454	220	4354	9	70	1	chevrolet impala
9	14	8	440	215	4312	8.5	70	1	plymouth fury iii
10	14	8	455	225	4425	10	70	1	pontiac catalina
11	15	8	390	190	3850	8.5	70	1	amc ambassador dpl
12	15	8	383	170	3563	10	70	1	dodge challenger se

Notes

1. The data is structured, consolidated, consistent and clean

The auto-mpg dataset

1	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
2	18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
3	15	8	350	165	3693	11.5	70	1	buick skylark 320
4	18	8	318	150	3436	11	70	1	plymouth satellite
5	16	8	304	150	3433	12	70	1	amc rebel sst
6	17	8	302	140	3449	10.5	70	1	ford torino
7	15	8	429	198	4341	10	70	1	ford galaxie 500
8	14	8	454	220	4354	9	70	1	chevrolet impala
9	14	8	440	215	4312	8.5	70	1	plymouth fury iii
10	14	8	455	225	4425	10	70	1	pontiac catalina
11	15	8	390	190	3850	8.5	70	1	amc ambassador dpl
12	15	8	383	170	3563	10	70	1	dodge challenger se

Notes

1. The data is structured, consolidated, consistent and clean
2. The first row contains the headings (column names), the remaining rows are the observations (cases, instances,. . .)

The auto-mpg dataset

1	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
2	18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
3	15	8	350	165	3693	11.5	70	1	buick skylark 320
4	18	8	318	150	3436	11	70	1	plymouth satellite
5	16	8	304	150	3433	12	70	1	amc rebel sst
6	17	8	302	140	3449	10.5	70	1	ford torino
7	15	8	429	198	4341	10	70	1	ford galaxie 500
8	14	8	454	220	4354	9	70	1	chevrolet impala
9	14	8	440	215	4312	8.5	70	1	plymouth fury iii
10	14	8	455	225	4425	10	70	1	pontiac catalina
11	15	8	390	190	3850	8.5	70	1	amc ambassador dpl
12	15	8	383	170	3563	10	70	1	dodge challenger se

Notes

1. The data is structured, consolidated, consistent and clean
2. The first row contains the headings (column names), the remaining rows are the observations (cases, instances,. . .)
3. Each column stores a *variable*. Other terms include attributes and targets.

The auto-mpg dataset

1	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
2	18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
3	15	8	350	165	3693	11.5	70	1	buick skylark 320
4	18	8	318	150	3436	11	70	1	plymouth satellite
5	16	8	304	150	3433	12	70	1	amc rebel sst
6	17	8	302	140	3449	10.5	70	1	ford torino
7	15	8	429	198	4341	10	70	1	ford galaxie 500
8	14	8	454	220	4354	9	70	1	chevrolet impala
9	14	8	440	215	4312	8.5	70	1	plymouth fury iii
10	14	8	455	225	4425	10	70	1	pontiac catalina
11	15	8	390	190	3850	8.5	70	1	amc ambassador dpl
12	15	8	383	170	3563	10	70	1	dodge challenger se

Notes

1. The data is structured, consolidated, consistent and clean
2. The first row contains the headings (column names), the remaining rows are the observations (cases, instances,. . .)
3. Each column stores a *variable*. Other terms include attributes and targets.
4. Machine learning uses combinations of these columns to build models.

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?

Applied to `auto-mpg`

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon...

Applied to `auto-mpg`

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon...
- But what collection of columns to use?

Applied to `auto-mpg`

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon…
- But what collection of columns to use?
- Depends on the problem we wish to solve…

Applied to `auto-mpg`

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon...
- But what collection of columns to use?
- Depends on the problem we wish to solve...
- Can it be used to predict some quantity (a target)?

Applied to `auto-mpg`

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon...
- But what collection of columns to use?
- Depends on the problem we wish to solve...
- Can it be used to predict some quantity (a target)?
- And what does *prediction* mean?

Applied to `auto-mpg`

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon…
- But what collection of columns to use?
- Depends on the problem we wish to solve…
- Can it be used to predict some quantity (a target)?
- And what does *prediction* mean?
- Are there other forms of learning apart from being able to predict?

Applied to `auto-mpg`

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon...
- But what collection of columns to use?
- Depends on the problem we wish to solve...
- Can it be used to predict some quantity (a target)?
- And what does *prediction* mean?
- Are there other forms of learning apart from being able to predict?

Applied to `auto-mpg`

- Given explanatory variables `displacement`, `horsepower`, `weight`, can we predict *mpg* (target)?

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon...
- But what collection of columns to use?
- Depends on the problem we wish to solve...
- Can it be used to predict some quantity (a target)?
- And what does *prediction* mean?
- Are there other forms of learning apart from being able to predict?

Applied to `auto-mpg`

- Given explanatory variables `displacement`, `horsepower`, `weight`, can we predict *mpg* (target)?
- Are all these explanatory variables needed, or could some be dropped?

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon...
- But what collection of columns to use?
- Depends on the problem we wish to solve...
- Can it be used to predict some quantity (a target)?
- And what does *prediction* mean?
- Are there other forms of learning apart from being able to predict?

Applied to `auto-mpg`

- Given explanatory variables `displacement`, `horsepower`, `weight`, can we predict *mpg* (target)?
- Are all these explanatory variables needed, or could some be dropped?
- Are additional explanatory variables needed, either from `auto-mpg` or elsewhere?

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon. . .
- But what collection of columns to use?
- Depends on the problem we wish to solve. . .
- Can it be used to predict some quantity (a target)?
- And what does *prediction* mean?
- Are there other forms of learning apart from being able to predict?

Applied to `auto-mpg`

- Given explanatory variables `displacement`, `horsepower`, `weight`, can we predict *mpg* (target)?
- Are all these explanatory variables needed, or could some be dropped?
- Are additional explanatory variables needed, either from `auto-mpg` or elsewhere?
- How do we measure the *quality* of a prediction?

Understanding the `auto-mpg` dataset: Column Sufficiency

Learning

- Given a collection of columns (a *projection* of the full dataset), how does this help the machine to learn?
- It provides example data representing a phenomenon…
- But what collection of columns to use?
- Depends on the problem we wish to solve…
- Can it be used to predict some quantity (a target)?
- And what does *prediction* mean?
- Are there other forms of learning apart from being able to predict?

Applied to `auto-mpg`

- Given explanatory variables `displacement`, `horsepower`, `weight`, can we predict *mpg* (target)?
- Are all these explanatory variables needed, or could some be dropped?
- Are additional explanatory variables needed, either from `auto-mpg` or elsewhere?
- How do we measure the *quality* of a prediction?
- What other learning can be derived from the chosen column collection?

Understanding the `auto-mpg` dataset: Row Sufficiency

Selection

- Do we have enough, too many or just enough observations?

Example

Understanding the `auto-mpg` dataset: Row Sufficiency

Selection

- Do we have enough, too many or just enough observations?
- If we project the data, we might have multiple rows with the same explanatory values but different target values...

Example

Understanding the `auto-mpg` dataset: Row Sufficiency

Selection

- Do we have enough, too many or just enough observations?
- If we project the data, we might have multiple rows with the same explanatory values but different target values…
- …Is this good or bad?

Example

Understanding the `auto-mpg` dataset: Row Sufficiency

Selection

- Do we have enough, too many or just enough observations?
- If we project the data, we might have multiple rows with the same explanatory values but different target values…
- …Is this good or bad?
- How can we exclude unnecessary or incompatible observations?

Example

Understanding the `auto-mpg` dataset: Row Sufficiency

Selection

- Do we have enough, too many or just enough observations?
- If we project the data, we might have multiple rows with the same explanatory values but different target values…
- …Is this good or bad?
- How can we exclude unnecessary or incompatible observations?
- …We can use *selection* (also known as *restriction*) - but how do we choose which rows to keep?

Example

```
SELECT displacement, horsepower, weight, mpg
FROM auto_mpg
WHERE horsepower > 79;
```

- SELECT clause: projection (restricting columns: column sufficiency)
- WHERE clause: selection (restricting rows: row sufficiency)

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to "see the wood for the trees"

- Generally, we want our data to be as granular as possible - more detail is *better*

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to "see the wood for the trees"

- Generally, we want our data to be as granular as possible - more detail is *better*
- ... We can remove detail if needed, but cannot add it later

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to "see the wood for the trees"

- Generally, we want our data to be as granular as possible - more detail is *better*
- . . . We can remove detail if needed, but cannot add it later
- Is duplicate data good or bad in machine learning?

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to “see the wood for the trees”

- Generally, we want our data to be as granular as possible - more detail is *better*
- ... We can remove detail if needed, but cannot add it later
- Is duplicate data good or bad in machine learning?
- ... GOOD: estimating variability in a quantity, using statistical methods

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to "see the wood for the trees"

- Generally, we want our data to be as granular as possible - more detail is *better*
- . . . We can remove detail if needed, but cannot add it later
- Is duplicate data good or bad in machine learning?
- . . . GOOD: estimating variability in a quantity, using statistical methods
- . . . BAD: can obscure useful implementation - an aggregated value might be more useful

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to "see the wood for the trees"

- Generally, we want our data to be as granular as possible - more detail is *better*
- ... We can remove detail if needed, but cannot add it later
- Is duplicate data good or bad in machine learning?
- ... GOOD: estimating variability in a quantity, using statistical methods
- ... BAD: can obscure useful implementation - an aggregated value might be more useful
- Can generate summaries in three ways

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to "see the wood for the trees"

- Generally, we want our data to be as granular as possible - more detail is *better*
- . . . We can remove detail if needed, but cannot add it later
- Is duplicate data good or bad in machine learning?
- . . . GOOD: estimating variability in a quantity, using statistical methods
- . . . BAD: can obscure useful implementation - an aggregated value might be more useful
- Can generate summaries in three ways
    1. Sampling - reduce/remove row duplication

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to "see the wood for the trees"

- Generally, we want our data to be as granular as possible - more detail is *better*
- ... We can remove detail if needed, but cannot add it later
- Is duplicate data good or bad in machine learning?
- ... GOOD: estimating variability in a quantity, using statistical methods
- ... BAD: can obscure useful implementation - an aggregated value might be more useful
- Can generate summaries in three ways
  1. Sampling - reduce/remove row duplication
  2. Banding - reduce the cardinality of a column

Understanding the `auto-mpg` dataset: Summarising

Often in data mining, we need to "see the wood for the trees"

- Generally, we want our data to be as granular as possible - more detail is *better*
- ...We can remove detail if needed, but cannot add it later
- Is duplicate data good or bad in machine learning?
- ...GOOD: estimating variability in a quantity, using statistical methods
- ...BAD: can obscure useful implementation - an aggregated value might be more useful
- Can generate summaries in three ways
    1. Sampling - reduce/remove row duplication
    2. Banding - reduce the cardinality of a column
    3. Grouped Aggregation - roll up by level, aggregating as needed

Understanding the `auto-mpg` dataset: Sampling

To reduce (but not remove) duplication, sampling can be a good compromise

- To reduce bias, the sample should be random (each row has the same probability of being picked)

Understanding the `auto-mpg` dataset: Sampling

To reduce (but not remove) duplication, sampling can be a good compromise

- To reduce bias, the sample should be random (each row has the same probability of being picked)
- Number of rows to keep in the sample is a compromise

Understanding the `auto-mpg` dataset: Sampling

To reduce (but not remove) duplication, sampling can be a good compromise

- To reduce bias, the sample should be random (each row has the same probability of being picked)
- Number of rows to keep in the sample is a compromise
- Can reduce runtime while allowing estimates of the uncertainty in a predictive model

Understanding the `auto-mpg` dataset: Sampling

To reduce (but not remove) duplication, sampling can be a good compromise

- To reduce bias, the sample should be random (each row has the same probability of being picked)
- Number of rows to keep in the sample is a compromise
- Can reduce runtime while allowing estimates of the uncertainty in a predictive model
- However, an aggregated column might be a better choice

```sql
-- Return a random sample of 3 Japanese cars
SELECT *
FROM AutoMpg
WHERE originID = 3
ORDER BY RANDOM()
LIMIT 3;
```

Understanding the `auto-mpg` dataset: Banding

A new column with reduced cardinality is often more understandable

- Sometimes a column with fewer distinct values offers fresh insights

Understanding the `auto-mpg` dataset: Banding

A new column with reduced cardinality is often more understandable

- Sometimes a column with fewer distinct values offers fresh insights
- . . . Derive `carMaker` from `carName` - compare different manufacturers

```
SELECT substr(carName,1,instr(carName,' ')-1) AS carMaker
 ,CASE
   WHEN horsepower < 130 THEN 'low'
   ELSE 'high'
  END AS horsepowerGroup
FROM AutoMpg;
```

Understanding the `auto-mpg` dataset: Banding

A new column with reduced cardinality is often more understandable

- Sometimes a column with fewer distinct values offers fresh insights
- ...Derive `carMaker` from `carName` - compare different manufacturers
- When a column contains real numbers, currency, etc., this is very noticeable

```sql
SELECT substr(carName,1,instr(carName,' ')-1) AS carMaker
 ,CASE
   WHEN horsepower < 130 THEN 'low'
   ELSE 'high'
  END AS horsepowerGroup
FROM AutoMpg;
```

Understanding the `auto-mpg` dataset: Banding

A new column with reduced cardinality is often more understandable

- Sometimes a column with fewer distinct values offers fresh insights
- . . . Derive `carMaker` from `carName` - compare different manufacturers
- When a column contains real numbers, currency, etc., this is very noticeable
- . . . *Banding* - assigning those numbers to non-overlapping ranges can simplify analysis

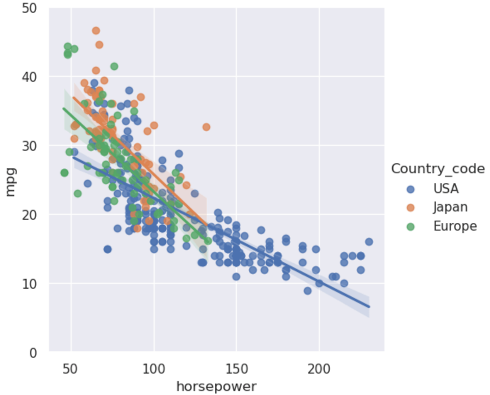
```
SELECT substr(carName,1,instr(carName,' ')-1) AS carMaker
 ,CASE
   WHEN horsepower < 130 THEN 'low'
   ELSE 'high'
  END AS horsepowerGroup
FROM AutoMpg;
```

Understanding the `auto-mpg` dataset: Banding

A new column with reduced cardinality is often more understandable

- Sometimes a column with fewer distinct values offers fresh insights
- . . . Derive `carMaker` from `carName` - compare different manufacturers
- When a column contains real numbers, currency, etc., this is very noticeable
- . . . *Banding* - assigning those numbers to non-overlapping ranges can simplify analysis

```
SELECT substr(carName,1,instr(carName,' ')-1) AS carMaker
 ,CASE
   WHEN horsepower < 130 THEN 'low'
   ELSE 'high'
  END AS horsepowerGroup
FROM AutoMpg;
```



Understanding the `auto-mpg` dataset: Grouped Aggregation

A grouped aggregation changes the effective key structure

- Aggregations include: `MIN()`, `SUM()`, `COUNT(DISTINCT ...)`

Understanding the `auto-mpg` dataset: Grouped Aggregation

A grouped aggregation changes the effective key structure

- Aggregations include: `MIN()`, `SUM()`, `COUNT(DISTINCT ...)`
- . . . Take a set of values, compute an aggregate value

Understanding the `auto-mpg` dataset: Grouped Aggregation

A grouped aggregation changes the effective key structure

- Aggregations include: `MIN()`, `SUM()`, `COUNT(DISTINCT ...)`
- ...Take a set of values, compute an aggregate value
- Sets can be partitioned by grouping variable, aggregate applied to each partition

Understanding the `auto-mpg` dataset: Grouped Aggregation

A grouped aggregation changes the effective key structure

- Aggregations include: `MIN()`, `SUM()`, `COUNT(DISTINCT ...)`
- ...Take a set of values, compute an aggregate value
- Sets can be partitioned by grouping variable, aggregate applied to each partition
- ...Example: average mpg per country of manufacture

```
SELECT originID, AVG(mpg)
FROM AutoMpg
GROUP BY originID;
```

Outline

Summary

- Semantically rich, flat data is preferred for machine learning
- Ideally, this data would also be structured, consolidated, consistent and clean
- Several data operations were described, using the AutoMpg dataset as an example data source
  - Projection
  - Selection
  - Summarising: Sampling, Banding, Grouped Aggregation

Your task is to apply this to datasets using the python toolchain.