CI330- Data Mining

University of Brighton |

Assignement 1

[Year]

Table of Contents

[Business Understanding: 2](#_Toc35429681)

[Business Intelligence: 2](#_Toc35429682)

[Legal & Ethical implications: 2](#_Toc35429683)

[Data Understanding & Preparation: 3](#_Toc35429684)

[Data Understanding & Preparation (continued): 4](#_Toc35429685)

[Data Understanding & Preparation (continued): 5](#_Toc35429686)

[Data Understanding & Preparation (continued): 6](#_Toc35429687)

[Data Understanding & Preparation (continued): 7](#_Toc35429688)

[Data Modelling: 8](#_Toc35429689)

[Star Schema: 8](#_Toc35429690)

[Data Mining Activity: 9](#_Toc35429691)

[Decision Tree: 9](#_Toc35429692)

[Overall View: 9](#_Toc35429693)

[0 to 1 Mile: 9](#_Toc35429694)

[1 – 2 miles: 10](#_Toc35429695)

[2 – 5 miles: 10](#_Toc35429696)

[5 – 10 miles: 11](#_Toc35429697)

[10+ miles: 11](#_Toc35429698)

[Naïve Bayes: 12](#_Toc35429699)

[Dependency Network: 12](#_Toc35429700)

[Attribute Characteristics: 12](#_Toc35429701)

[Number of children at home: 12](#_Toc35429702)

[Number of cars owned: 13](#_Toc35429703)

[Occupation: 13](#_Toc35429704)

[Attribute Profiles: 14](#_Toc35429705)

[Occupation: 14](#_Toc35429706)

[Number of children at home: 14](#_Toc35429707)

[Number cars owned: 15](#_Toc35429708)

[Attribute Discrimination: 15](#_Toc35429709)

[Clustering: 15](#_Toc35429710)

[Clustering Diagram: 15](#_Toc35429711)

[Cluster 3 tells us the following: 15](#_Toc35429712)

[Cluster 5 tells us the following: 15](#_Toc35429713)

[Cluster Profiles: 16](#_Toc35429714)

[Evaluation: 16](#_Toc35429715)

# Business Understanding:

In this assignment I will be investigating the commute distance of various customers in the Adventure Works database as a functional area of the business. Furthermore, researching its impact on various data; such as, commute distance in correlation to the number of cars owned or number of children at home. The main business objective of this investigation is the impact a person’s commute distance has.

## Business Intelligence:

If a business were to adopt the below approach of data mining activities and implement them into their business and use the gained knowledge, they could identify customers patterns and the relation of various data. For instance, in this report I investigate the relationship of a customers commute distance and number of children at home, cars owned, job description and many others.

## Legal & Ethical implications:

If the company decided to give customers the option of opting out of their data being collected this could remove the companies competitive edge over other companies, however, if this option is not given then there is the ethical dilemma of collection a customer’s information without their consent or knowledge. With the current laws and regulations set out by GDPR this process has become more streamlined, clear cut and transparent. However, once again there is the issue of companies (especially extremely large ones) not fully complying. For instance, when a customer’s account is created and they make a purchase, data such as commute distance, occupation, number of cars owned, etc… is collected and can be used for various activities. In this instance to find the relation between a customer commute distance and other data. If the customer is not made aware of this and or given the option to opt out this is none compliant which raises both legal & ethical issues. Another issue that must be taken into consideration is the company must ensure that their data is integral to ensure that all data mining activities and tools are accurate. For example, domain names, data types, sizes and acceptable ranges.

# Data Understanding & Preparation:

|  |  |  |
| --- | --- | --- |
| **Data in Model:** DimCustomer | **Source of data:** Adventure Works Data Dictionary | **Transformation:** |
| CustomerKey | Internal  AdventureWorks.Sales.CustomerID | Unique key for customer generated upon creation of a customer account.  PK\_Customer\_CustomerID  Primary Key (clustered) constraint  Joined to Sales.SalesOrderHeader.CustomerID |
| GeographyKey | Internal | Unique key generated upon creation of customer |
| CustomerAlternateKey | Internal | Generated upon creation of customer |
| Title | Internal  AdventureWorks.Person.Person.title | Joined to customer.  Example data: Mr., Mrs., Ms, Dr, Mx, etc…  SELECT Title  FROM person.person |
| FirstName | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vSalesPerson  AdventureWorks.Sales.vIndividualCustomer  AdventureWorks.Person.vAdditionalContactInfo | First name of the customer.  SELECT FirstName  FROM person.person |
| LastName | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vSalesPerson  AdventureWorks.Sales.vIndividualCustomer  AdventureWorks.Person.vAdditionalContactInfo | Last name of customer.  SELECT LastName  FROM person.person |

## Data Understanding & Preparation (continued):

|  |  |  |
| --- | --- | --- |
| **Data in Model:** DimCustomer | **Source of data:** Adventure Works Data Dictionary | **Transformation:** |
| NameStyle | Internal  AdventureWorks.Person.Person | Data stored in eastern style; i.e. First Name + Last Name.  e.g. John Smith |
| Birthdate | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics | Date of birth of customer.  Stored as datetime:  YYYY-MM-DD |
| MaritalStatus | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics | Data stored as nchar(1):  M = married  S = single |
| Suffix | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vIndividualCustomer | Suffix from surname  Data stored example: Jr. or Sr. |
| Gender | Internal  AdventureWorks.Sales.vPersonDemographics.Gender | Data stored as:  M = male  F = female  Neutral titles such as Dr, will produce issues; workaround, if one gender more probable, become default value. |

## Data Understanding & Preparation (continued):

|  |  |  |
| --- | --- | --- |
| **Data in Model:** DimCustomer | **Source of data:** Adventure Works Data Dictionary | **Transformation:** |
| EmailAddress | Internal  AdventureWorks.Person.EmailAddress  AdventureWorks.Production.ProductReview  AdventureWorks.Sales.vIndividualCustomer | Email address of customer.  Primary key with auto increment.  person join email address join businessEntityID  person join person join businessEntityID  PK\_EmailAddress\_BusinessEntityID\_EmailAddressID holds a primary key (clustered) constraint. |
| YearlyIncome | External  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics.YearlyIncome | Based on demographic of person join person table.  Source of data:   * gained through questionnaire * average income based on area from Geography Key. * use prediction algorithms based off already accumulated data, e.g. age, location, gender, position. * based on purchase patterns.   person join person  sales join vPersonDemographics |
| TotalChildren | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics | joined to person join person  SELECT TotalChildren  FROM Sales.vPersonDemographics |
| NumberChildrenAtHome | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics | joined to person join person |

## Data Understanding & Preparation (continued):

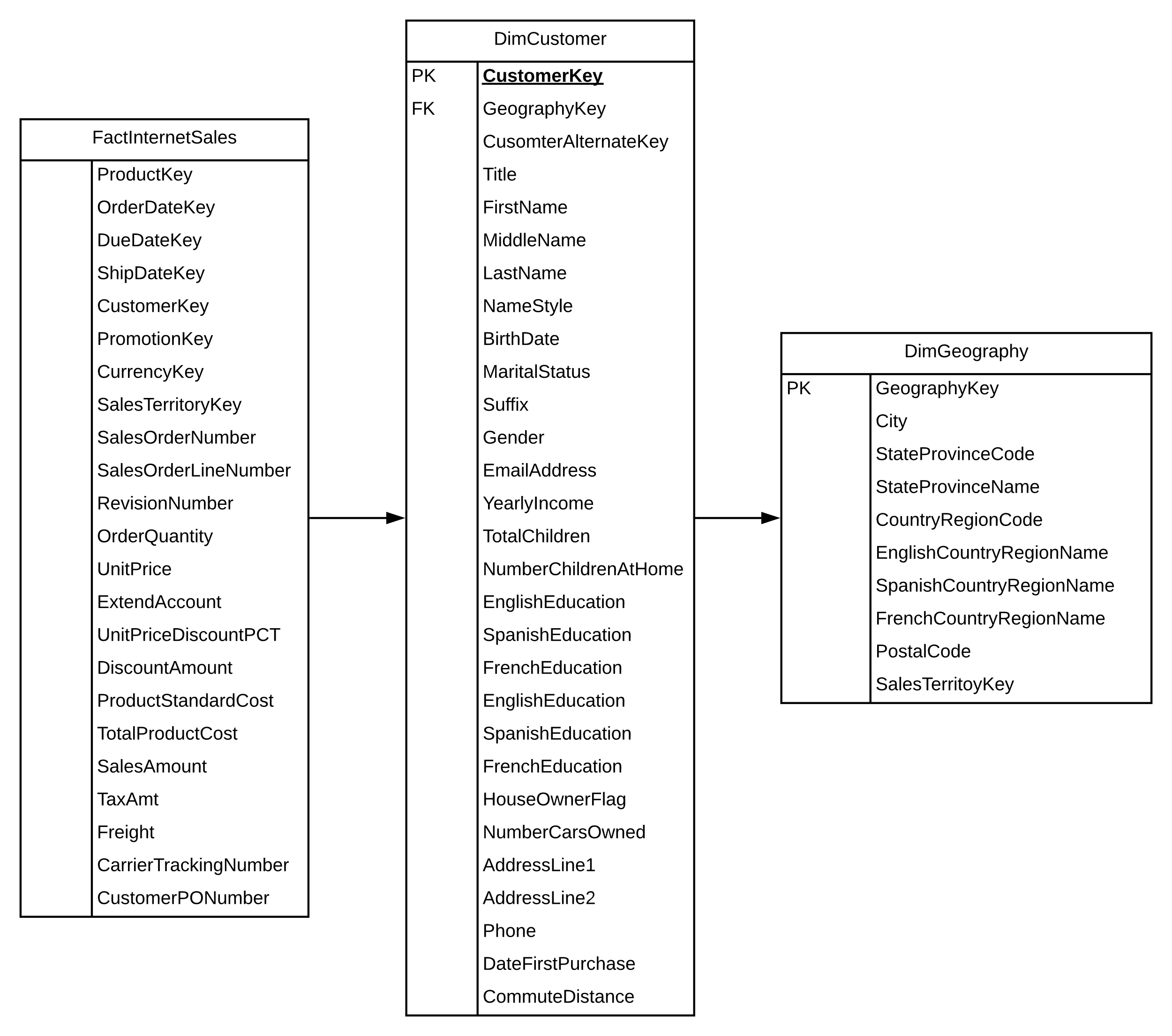
|  |  |  |
| --- | --- | --- |
| **Data in Model:** DimCustomer | **Source of data:** Adventure Works Data Dictionary | **Transformation:** |
| EnglishEducation | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics | joined person join person  sales join vPersonDemographics |
| SpanishEducation | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics | joined person join person  sales join vPersonDemographics |
| FrenchEducation | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics | joined person join person  sales join vPersonDemographics |
| NumberCarsOwned | Internal  AdventureWorks.Person.Person  AdventureWorks.Sales.vPersonDemographics | joined to person join person |
| AddressLine1 | Internal  AdventureWorks.Person.Address  AdventureWorks.Sales.vIndividualCustomer | First line of street address,  Joined to Customer join person join business entity join address |
| AddressLine2 | Internal  AdventureWorks.Person.Address  AdventureWorks.Sales.vIndividualCustomer | Second line of street address  Joined to Customer join person join business entity join address |

## Data Understanding & Preparation (continued):

|  |  |  |
| --- | --- | --- |
| **Data in Model:** DimCustomer | **Source of data:** Adventure Works Data Dictionary | **Transformation:** |
| Phone | Internal  AdventureWorks.Person.PersonPhone  AdventureWorks.Person.PersonPhone.PhoneNumber  AdventureWorks.Person.PersonPhone. PhoneNumberTypeID  AdventureWorks.Sales.vIndividualCustomer.PhoneNumber  AdventureWorks.Sales.vIndividualCustomer.PhoneNumberType | FK\_PersonPhone\_Person\_BusinessEntityID  foreign key constraint referencing Person.BuinessEntityID.  Person.Person.BusinessEntityID join Person.PersonPhone.BusinessEntityID  Person.Personphone.BusinesssEntityID join Person.PersonBusinessEntityID |
| DateFirstPurchase | Internal  AdventureWorks.Sales.vPersonDemographics  AdventureWorks.Sales.SalesOrderHeader.OrderDate | Person.Person join Sales.vPersonDemographics |
| CommuteDistance | Internal |  |

# Data Modelling:

## Star Schema:



# Data Mining Activity:

## Decision Tree:

### Overall View:

|  |  |
| --- | --- |
|  | When investigating the relation between commute distance and customer using a decision tree algorithm, I found that commuters had the following percentage of commute distance:   * 37.18% between 0 and 1 mile. * 17.42% between 1 and 2 miles. * 17.6% between 2 to 5 miles. * 17.3% between 5 and 10 miles. * 13.49% more than 10 miles.   Bellow I explore this data further. |

### 0 to 1 Mile:

When looking at a commute distance between 0 and 1 mile the most common customer owned 0 cars and held a management position. The data further tells us that for commuters who owned 0 cars and held a management position had the following results:

|  |  |  |
| --- | --- | --- |
| **Number of children at home:** | **Income:** | **Cases:** |
| 0 children at home | between 74,000 and 988,888 | 37 |
| 0 children at home | more than or equal to 988,888 | 28 |
| 1 or more children at home | between 11,000 and 74,000 | 29 |

For commuters owning 0 cars and holding a professional job with 0 children at home the most common were those who had an income between 74,000 and 90,000 (94 cases) or commuters earning bellow 42,000 (85 cases). Moreover, commuter falling into the manual worker category owning 0 cars the most common type of customer was those who had 2 or more children (142 cases) or 0 children (243 cases). When looking at those holding a cleraical job with 0 children income was the most important factor, those who earned more than 26,000 held 644 cases and those who earned less than 26,000 held 271 cases.

For customers whose commute fell in the category of 0 to 1 mile the most likely customers that own 1 car would be as follows. For those who are manual workers the most likely customers were the ones who had 2 children at home (83 cases) and 3 children at home (50 cases). However, if a manual worker had 4 children their income would have to be less than 26,000 (52 cases).

For those commuters that owned 2 cars the majority of customers were manual workers who had more than 1 child (174 cases). The next highest type of customer are skilled manual workers who have 1 child at home and have a yearly income between 32,400 and 42,000 (71) cases. Furthermore, workers who held a management position and earned up to or less than 122,000 had 54 cases.

Furthermore, when we look at customers who own 3 cars, the most prominent were commuters who had 5 children, an income between 74,000 and 90,000 (52 cases). For those that had 4 children and an income equal to or more than 58,000 we found out they were skilled manual workers (13 cases). If the customer had one child however their income would have to be less than 122,000 (21 cases).

### 1 – 2 miles:

When looking at a commute distance between 1 and 2 miles the most commong customers were as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N’ Cars owned:** | **Job title:** | **Income:** | **Children at home:** | **N’ Cases:** |
| 2 | Management | Between 74,000 and 122,000 | 4 | 40 |
| 4 | N/A | Between 74,000 and 90,000 | 5 | 38 |
| 3 | Professional | Between 10,600 and 74,000 | 0 | 19 |
| 3 | Professional | Less than 42,000 | 0 | 7 |

### 2 – 5 miles:

When looking at the data bellow it is easy to assume the most likely customer is a commuter who owns 2 cars, if a professional and earns between 42,000 and 74,000 as this has the greatest number of cases. Though this is true one most also consider how many these cases represent of the total amount. For instance; a professional who owns 2 cars and earns between 74,000 and 90,000 only has 22 cases but those cases are 22 out of 34. That’s 64.7% of the total commuters who fall under that category.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N’ Cars owned:** | **Job title** | **Income:** | **Children at home:** | **N’ Cases:** |
| 1 car | Manual | Less than 26,000 | <2 or >4 | 348 |
| 1 car | Manual | >= 26,000 | <2 or >3 | 103 |
| 1 car | Manual |  |  |  |
|  |  |  |  |  |
| 1 | Skilled manual | >=42,000 | >=1 | 107 |
| 1 | Skilled manual | >=42,000 | 0 | 178 |
| 2 | Skilled manual | N/A | 2 to 3 | 118 |
| 0 | Skilled manual | N/A | 2 | 19 |
|  |  |  |  |  |
| 2 | Professional | Between 74,000 and 90,000 | 3 | 22 |
| 2 | Professional | Between 42,000 and 74,000 | 0 | 293 |
| 2 | Professional | >= 90,000 | 0 | 23 |

### 5 – 10 miles:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N’ Cars owned:** | **Job title** | **Income:** | **Children at home:** | **N’ Cases:** |
| 0 | Management | Between 74,000 and 110,000 | 1 or more | 13 / 25 |
| 2 | Management | Between 74,000 and 122,000 | 4 children | 14 |
| 2 | Management | Less than 42,000 | N/A | 62 / 110 |
|  |  |  |  |  |
| 0 | Professional | N/A | 1 | 27 / 46 |
|  |  |  |  |  |
| 2 | Skilled manual | Between 54,800 and 61,200 | 0 | 526 |
|  |  |  |  |  |
| 3 | N/A | Between 10,600 and 122,000 | 4 | 43 / 46 |
| 3 | N/A | >= 122,000 | 4 | 19 |
| 3 | N/A | Between 90,000 and 122,000 | 3 | 40 |
| 3 | N/A | Less than 42,000 | 3 | 17 |

### 10+ miles:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N’ Cars owned:** | **Job title** | **Income:** | **Children at home:** | **N’ Cases:** |
| 3 | N/A | Between 9,000 and 42,000 | 3 | 30 |
| 3 | N/A | Less than 42000 | 3 | 17 |
| 3 | N/A | Between 10,600 and 122,000 | 5 | 77 |
| 3 | N/A | Between 74,000 and 61,200 | 5 | 78 |
|  |  |  |  |  |
| 3 | Management | Between 10,600 and 90,000 | 5 | 38 |
|  |  |  |  |  |
| 3 | Professional | Between 66,888 and 74,000 | 0 | 56 |
| 3 | Professional | Less than 66,888 | 0 | 18 / 29 |
| 3 | Professional | >= 74,000 | 0 | 157 |
| 4 | Professional | >=74,000 | 0 | 57 |
| 4 | Professional | Between 61,200 and 74,000 | 0 | 20 |
| 2 | Professional | >=90,000 | 5 | 22 / 34 |
|  |  |  |  |  |
|  |  |  |  |  |
| 3 | Skilled manual | <58,000 | 4 | 17/18 |
| 2 | Skilled manual | >=42,000 | 1 | 29/37 |

## Naïve Bayes:

### Dependency Network:

The dependency network shows that when inspecting commute distance 5 factors come into play with different levels of importance respectively, number of children at home, occupation (Spanish, French & English) and number of cars owned. The strongest link for commute distance is the number of cars owned.

### Attribute Characteristics:

#### Number of children at home:

From the graph below it is extremely apparent that the most common commuter has 0 children at home and has a commute distance between 1 and 2 miles.

#### Number of cars owned:

From the graph below it is easy to tell the most common commuter owns 2 cars and has a commute between 5 and 10 miles.

#### Occupation:

The graph below tells us that the most common commuter has a “professional” profession and has a commute distance greater than 10 miles.

### Attribute Profiles:

#### Occupation:

#### Number of children at home:

#### Number cars owned:

### Attribute Discrimination:

When comparing a commute distance of 0 to 1-mile vs a 2 – 5-mile commute it becomes apparent that French, Spanish & English occupations such as clerical, manual and professional will favour a 0 to 1-mile commute. Moreover, especially if they do not own a car and have no children. However, if 1 car is owned and or the they hold a professional occupation there is a clear favour towards the 2 – 5-mile commute. If 2 cars are owned there is also a favour towards a 1 to 2-mile commute and if they are skilled manual. Professional and management occupations appear to favour a 10+ mile commute.

## Clustering:

### Clustering Diagram:

The clustering diagram shows us that the strongest link is between cluster 5 and cluster 3.

#### Cluster 3 tells us the following:

|  |  |
| --- | --- |
| **Variable:** | **Values:** |
| French occupation | cadre |
| Spanish occupation | professional |
| English occupation | professional |
| number of children at home | 0 |
| yearly income | 57,294.7 – 79,108 |
| number cars owned | 1 |

#### Cluster 5 tells us the following:

|  |  |
| --- | --- |
| **Variable:** | **Values:** |
| French occupation | cadre |
| Spanish occupation | professional |
| English occupation | professional |
| number of cars owned | 2 |
| number of children at home | 0 |
| yearly income | 57,294.7 – 79,108 |

### Cluster Profiles:

The highest probably were commuters who had 0 children at home, owned 2 cars, had a commute distance between 0 and 1 mile.