

# Notes for the BAN400 Exam

## Table of contents

<b>1 Functions</b>	<b>1</b>
1.1 Basic functions . . . . .	1
1.2 Math . . . . .	2
1.3 Reading data . . . . .	2
1.4 Data wrangling . . . . .	2
1.5 Machine learning . . . . .	3
1.6 Many models . . . . .	3
1.7 Parsing . . . . .	3
1.8 Selecting . . . . .	4
1.9 Parallel computing . . . . .	4
1.10 Making maps . . . . .	4
<b>2 Topics</b>	<b>4</b>
2.1 Filtering . . . . .	4
2.2 Selecting . . . . .	5
2.3 Loops and iterations . . . . .	5
2.4 Math . . . . .	6
2.5 Plotting . . . . .	6
2.6 Reading data . . . . .	8
2.7 Parsing data . . . . .	8

## 1 Functions

### 1.1 Basic functions

Function	Package	Description
mean()	base	Calculates the mean of a vector of numbers
median()	base	Calculates the median of a vector of numbers
sd()	base	Calculates the standard deviation of a vector of numbers
var()	base	Calculates the variance of a vector of numbers
sum()	base	Calculates the sum of a vector of numbers
c()	base	Creates vector
length()	base	The number of elements in a vector or list
ncol()	base	Number of columns of data frame or matrix
nrow()	base	Number of rows of data frame or matrix
min()	base	The smallest value in a set
max()	base	The largest value in a set

## 1.2 Math

Function	Package	Description
<code>sqrt()</code>	base	Calculates square root of number or vector of numbers
<code>abs()</code>	base	Calculates absolute value of number or vector of numbers
<code>sign()</code>	base	Returns the sign of x
<code>round()</code>	base	Rounds x to n decimal places
<code>ceiling()</code>	base	Rounds up to the nearest integer
<code>floor()</code>	base	Rounds down to the nearest integer
<code>cumsum()</code>	base	Cumulative sum
<code>cor()</code>	base	Correlation

## 1.3 Reading data

Function	Package	Description
<code>read_delim()</code>	readr	Read file with columns separated by any delimiter
<code>read_csv()</code>	readr	Read csv-file (comma separated values)
<code>read_csv2()</code>	readr	Uses semicolons as separators and keeps commas
<code>read_excel()</code>	readxl	Read data from excel files
<code>read_fwf()</code>	readr	Reads fixed width data
<code>read_tsv()</code>	readr	Reads tab separated values

## 1.4 Data wrangling

Function	Package	Description
<code>head()</code>	base	Returns the first few rows of a data frame or vector
<code>tail()</code>	base	Returns the first few rows of a data frame or vector
<code>filter()</code>	dplyr	Returns elements that satisfy conditions
<code>select()</code>	dplyr	Choose specific columns from a data frame
<code>arrange()</code>	dplyr	Sorts rows of a data frame by specified columns
<code>sort()</code>	base	Sorts a vector in ascending or descending order
<code>mutate()</code>	dplyr	Adds or modifies columns in a data frame
<code>transmute()</code>	dplyr	Creates a new data frame containing only the specified computations
<code>summarise()</code>	dplyr	Summary statistics for columns in a data frame (typically used with grouped data)
<code>group_by()</code>	dplyr	Group data by one or more variables
<code>ungroup()</code>	dplyr	Ungroup data such that subsequent operations to apply to the entire dataset
<code>left_join()</code>	dplyr	Returns all values from the first data frame with all columns and values from the second data frame where there is a match
<code>inner_join()</code>	dplyr	Joins two data frames by keeping only records that match in both data sets
<code>right_join()</code>	dplyr	Returns all values from the second data frame with matching columns and values from the first data frame where there is a match
<code>full_join()</code>	dplyr	Returns all values and columns from both data frames and filling in NA where there is no match
<code>semi_join()</code>	dplyr	Filters the first data frame keeping only rows with matching keys in the second data frame
<code>anti_join()</code>	dplyr	Filters the first data frame to keep only rows with no match in the second data frame

## 1.5 Machine learning

Function	Package	Description
logistic_reg()	tidymodels	Specifies a logistic regression model
set_engine()	tidymodels	Specifies the computational engine for a model
set_mode()	tidymodels	Sets the mode (e.g. classification or regression) for a model
fit()	tidymodels	Fits the model to data
nearest_neighbor()	tidymodels	Specifies a k-nearest neighbors model
tune()	tidymodels	Marks a parameter for tuning in a model
finalize_workflow()	tidymodels	Finalizes the workflow with specific parameters
workflow()	tidymodels	Creates a workflow object
add_model()	tidymodels	Adds a model to a workflow
add_recipe()	tidymodels	Adds a recipe to a workflow
tune_grid()	tidymodels	Tunes hyperparameters across a grid of values
select_best()	tidymodels	Selects the best tuning parameter combination based on a metric

## 1.6 Many models

Function	Package	Description
add_predictions()	modelr	Adds model predictions to a data frame
add_residuals()	modelr	Adds residuals from a model to a data frame
group_by()	dplyr	Groups data by one or more variables
ungroup()	dplyr	Removes grouping structure from a data frame
nest()	tidyr	Creates a nested data frame by collapsing rows into list-columns
unnest()	tidyr	Expands list-columns back into regular columns
select()	dplyr	Selects specific columns from a data frame
pull()	dplyr	Extracts a single column as a vector
pluck()	purrr	Extracts an element from a list or vector by index or name
map()	purrr	Applies a function to each element of a list or vector
map2()	purrr	Applies a function to pairs of elements from two lists
glance()	broom	Generates a summary of model diagnostics in a tidy format

## 1.7 Parsing

Function	Package	Description
class()	base	Returns the class of an object
typeof()	base	Determines the type or storage mode of any object
mode()	base	Indicates the mode of storage
as.numeric()	base	Converts data to numeric format
as.character()	base	Converts data to character format
as.factor()	base	Converts data to factor format
gsub()	base	Find and replace
reshape()	base	Reshape datasets between wide and long formats
pivot_longer()	tidyr	Convert wide data to long format
pivot_wider()	tidyr	Convert long data to wide format.
na.omit()	base	Remove missing values
is.na()	base	Check for missing values
ymd()	lubridate	Parse dates
hms()	lubridate	Parse times
parse_datetime()	readr	Parse datetimes
separate()	tidyr	Split one column into multiple columns

## 1.8 Selecting

Function	Package	Description
<code>any_of()</code>	<code>dplyr</code>	Selects columns that match any of the given names in a vector
<code>all_of()</code>	<code>dplyr</code>	Selects columns that match all the given names in a vector
<code>starts_with()</code>	<code>dplyr</code>	Selects columns whose names start with a specified prefix
<code>ends_with()</code>	<code>dplyr</code>	Selects columns whose names end with a specified suffix
<code>contains()</code>	<code>dplyr</code>	Selects columns whose names contain a specified string
<code>matches()</code>	<code>dplyr</code>	Selects columns whose names match a specified regular expression
<code>select()</code>	<code>dplyr</code>	Selects specified columns from a data frame
<code>slice_min()</code>	<code>dplyr</code>	Selects rows with the smallest values of a variable
<code>slice_max()</code>	<code>dplyr</code>	Selects rows with the largest values of a variable

## 1.9 Parallel computing

Function	Package	Description
<code>tic()</code>	<code>tictoc</code>	Initialize time taking of a function
<code>toc()</code>	<code>tictoc</code>	Stop time taking of a function
<code>detectCores()</code>	<code>doParallel</code>	Detect cores in CPU
<code>makeCluster()</code>	<code>doParallel</code>	Initialize cores
<code>registerDoParallel</code>	<code>doParallel</code>	Register the cluster
<code>stopCluster()</code>	<code>doParallel</code>	Close off clusters
<code>foreach() %dopar%</code>	<code>doParallel</code>	Designate tasks for parallelization

## 1.10 Making maps

Function	Package	Description
<code>distm()</code>	<code>geosphere</code>	Calculate distance between 2 points
<code>st_as_sf()</code>	<code>sf</code>	Converting data frame to geometric object
<code>geom_sf()</code>	<code>sf</code> & <code>ggplot2</code>	Allows mapping of geometric objects
<code>st_crop()</code>	<code>sf</code>	Cropping maps (x and y chords)
<code>st_layers()</code>	<code>sf</code>	Check layers of files
<code>st_read()</code>	<code>sf</code>	Read map data
<code>st_transform()</code>	<code>sf</code>	Transform map projection

## 2 Topics

### 2.1 Filtering

The symbol `|` works as an “or” operator, meaning it will return items that satisfy either one or both of the conditions before and after the operator.

```
filter(flights, dest == "IAH" | dest == "HOU")
```

The `&` is an “and” operator, meaning both conditions need to hold. This is useful in conjunction with the `|`-operator, since otherwise you can simply just add filters after each other like this:

```
filter(!is.na(flight2), !is.na(flight1))
```

Another operator `%in%` returns `TRUE` if an item exists inside a vector:

```
flights %>%
  filter(dest %in% c("IAH", "HOU"))
```

To create the opposite logic, i.e. “not in c()”, you can use !:

```
flights %>%
  filter(!(dest %in% c("IAH", "HOU")))
```

## 2.2 Selecting

Some basic selection functions from dplyr are `starts_with`, `ends_with` and `contains`.

```
select(flights, starts_with("dep_"), starts_with("arr_"))
```

### 2.2.1 Regex (regular expressions)

The function `matches` uses *regular expressions*. These can be used to match precise patterns.

```
select(flights, matches("^(dep|arr)_(time|delay)$"))
```

Some common Regex meta characters are:

- `.` Matches any character except newline.
- `^` Matches the beginning of a string.
- `$` Matches the end of a string.
- `[]` Matches any one of the characters inside the brackets.
- `|` Logical OR.
- `*` Matches 0 or more repetitions of the preceding character.
- `+` Matches 1 or more repetitions.
- `?` Matches 0 or 1 repetition (optional match).
- `{n,m}` Matches between n and m repetitions.

## 2.3 Loops and iterations

### 2.3.1 Standard for-loop

```
for(i in 1:n) {
  ... do something with i...
}
```

Note that we can iterate over any type of vector, not just numbers, and we can give the iteration variable any name we want. In the example above it is `i`.

### 2.3.2 While loop

Repeat until a certain condition is met. For example

```
i <- 1
while(i < 10) {
  print(i)
  i <- i + 1
}
```

## 2.4 Math

The Integer Division Operator `%%` performs integer division. It divides two numbers and returns the whole number part of the quotient, discarding any remainder. Example:

```
10 %% 3 # Returns 3 (quotient without remainder)
```

The Modulo Operator `%%` operator returns the remainder from the division of two numbers. Example:

```
10 %% 3 # Returns 1 (remainder)
```

## 2.5 Plotting

We use `ggplot2` as the standard package for plotting, and the main function is `ggplot`. We supply a data frame to the first argument and an aesthetic mapping to the second argument. We add layers of plotting components using the plus sign. A simple example:

```
ggplot(df, aes(x = x_variable, y = y_variable, colour = grouping_variable)) +  
  geom_point()
```

Many types of layers may contain other data sets via the `data` argument and/or updated aesthetic mappings via the `mapping` argument. Data and mappings are typically inherited from the layer above if not specified in a new layer. There are many types of functions for making further adjustments to labels, titles, axes and other properties. A more complete example may look like this:

```
ggplot(df, aes(x = x_variable, y = y_variable)) +  
  geom_point() +  
  geom_point(mapping = aes(x = new_x_variable, y = new_y_variable, colour = grouping_var),  
    data = new_df) +  
  xlab("X label") +  
  ylab("Y label") +  
  ggtitle("Title of plot") +  
  theme_minimal()
```

### 2.5.1 Grouping

You can visually group items in plots together using the `color` argument.

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color=drv))
```

Grouping can also be used in calculations. For example, if you want to create several regression lines based on classification, you can group with `group` argument:

```
geom_smooth(aes(group = drv)) # Creates separate regression lines.
```

Sometimes need to group by several variables at the same time. You can use `unite()` to achieve this.

```
whodata %>%  
  unite(country_sex, country, sex, remove = FALSE) %>%  
  ggplot()+  
  geom_line(aes(x=year,y=cases, group=country_sex, color=sex))
```

### 2.5.2 Stacking plots

You can stack several plots on top of each other. If you stack plots of the same type, it may be useful to distinguish them using color. For example:

```
ggplot(diamonds) +  
  geom_freqpoly(aes(x = x, color = "x"), binwidth = 0.1) +  
  geom_freqpoly(aes(x = y, color = "y"), binwidth = 0.1)
```

### 2.5.3 One categorical and one continuous variable

To explore the relationship between a continuous variable and one categorical variable, some plotting options are:

- `geom_violin`
- `geom_freqpoly` with `colour`-argument
- `geom_histogram` with `facet_wrap(vars())`
- If the data set is large, you may want to use `geom_lv` from the library `lvplot`.

Example:

```
ggplot(diamonds, aes(x=price, colour=cut))+  
  geom_freqpoly(bins=50)
```

### 2.5.4 Two categorical variables

To explore the relationship between two categorical variables, `geom_tile` is often useful.

```
diamonds %>% count(cut,color) %>% ggplot(aes(x=cut,y=color))+  
  geom_tile(aes(fill=n))
```

If you want to normalize proportions within a given categorical variable, here's an example of that:

```
diamonds %>%  
  count(color, cut) %>%  
  group_by(color) %>%  
  mutate(  
    prop=n/sum(n)  
  ) %>%  
  ggplot(aes(x=color,y=cut))+  
  geom_tile(aes(fill=prop))
```

### 2.5.5 Two continuous variables

There are many options to explore the relationship between two categorical variables, such as:

- `geom_point`
- `geom_smooth`
- `geom_boxplot` where a continuous variable is divided into bins.

```
ggplot(diamonds, aes(x = cut_width(price, 2000, boundary = 0), y = carat)) +  
  geom_boxplot(varwidth = TRUE)
```

**Tip:** choosing binned plots versus such as scatterplots depends on the nature of data analysis. For example, binned plots can “hide” outliers that are unusual combinations of x and y, but where x or y individually are not extreme values.

### 2.5.6 Discrete variables

A small tip about plotting discrete variables: If there is overplotting (e.g. multiple “dots” overlapping in dotplots; which happens often with discrete variables), then use jitter functions.

```
ggplot(data, aes(x = category, y = value)) +  
  geom_jitter(width = 0.2, height = 0.1)
```

### 2.5.7 Statistics

Most geoms come in pairs with complementary statistics arguments that are almost always used in concert. These functions can be used to retrieve the data that is used to generate the plot. For example, for these geoms:

```
geom_smooth()
geom_dotplot()
geom_point()
geom_bar()
```

We have these stat functions respectively:

```
stat_smooth()
stat_bindot()
stat_qq()
stat_count()
```

The corresponding stat functions can be found by reading the documentation with for example `?geom_smooth`.

## 2.6 Reading data

Sometimes data contains the delimiter. In that case, use quote argument to escape:

```
read_csv("x,y\n1,'a,b'", quote="''")
```

## 2.7 Parsing data

There are various parsing functions from `lubridate` library, such as `ymd()`. These functions can be used to convert objects into time and date formats.

### 2.7.1 Dates and time

Examples:

```
d1 <- "January 1, 2010"
d2 <- "2015-Mar-07"
d3 <- "06-Jun-2017"
d4 <- c("August 19 (2015)", "July 1 (2015)")
d5 <- "12/30/14" # Dec 30, 2014
t1 <- "1705"
t2 <- "11:15:10.12 PM"
```

Can be parsed respectively:

```
mdy(d1)      #> [1] "2010-01-01"
ymd(d2)      #> [1] "2015-03-07"
dmy(d3)      #> [1] "2017-06-06"
mdy(gsub("\\(", "", gsub("\\)", "", d4))) #> [1] "2015-08-19" "2015-07-01"
mdy(d5)      #> [1] "2014-12-30"
hm(t1)       #> [1] "17:05:00"
hms(t2)      #> [1] "23:15:10.12"
```



### 2.7.2 Pivoting

The library `tidyr` provides various ways to pivot data. An example of a pivot is from

```
preg <- tribble(
  ~pregnant, ~male, ~female,
  "yes", NA, 10,
  "no", 20, 12
)
```

To a longer data format:

```
preg_tidy2 <- preg %>%
  pivot_longer(c(male, female), names_to = "sex", values_to = "count", values_drop_na = TRUE)
```

### 2.7.3 Splitting columns

You can split a column into multiple columns using `separate()` and `extract()` from `tidyr`:

```
tibble(x = c("a,b,c", "d,e,f,g", "h,i,j")) %>%
  separate(x, c("one", "two", "three"))
```

In the above case, the second item has four “values” “d,e,f,g”. In this case, the `g` is dropped. There are various arguments in the functions to determine how to deal with extra or missing items like this.

### 2.7.4 Basic find and replace

You can perform a basic find-and-replace operation on a vector like this:

```
y[y == "find"] <- "replace"
```