CA Instructions

- Add your student ID to the set.seed() function, and run the starting code.
- Do all foundations questions.
- Do at least 4 other questions (your best 4 will be used to determine your score).
- Copy+paste the output noting the columns you lost.
- For each question, provide a screen shot of the output, and a brief discussion on the approach taken in a word document (upload this to moodle).
- Prepare an executable solution for me to be able to recreate your answers (upload this to moodle).

Foundations Questions – up to 7 points (10 including the optional show-off question).

[F-1] – Ensure that each attribute is appropriately encoded, i.e. as a character, numeric, factor (with correct levels).
[F-2] – Identify which attributes have missing values, and how many.
[F-3] – Deal with an attribute that has missing values (show the outcome of your approach by redoing F-2). Options to do this are:

[Basic] – removing the column
[Intermediate] – assigning the mean/median or mode (if categorical) value, or
[Advanced] – imputing the missing values

OPTIONAL Show-Off Foundations Question
[F-4] Check for outliers; if you find some, decide if you want to keep them in the dataset or not (defend your choice!) – you don't need to remove them though, just show how you would find them and discuss their presence.

Hint: some advanced/show-off questions can't deal with missing values. Once you've done F3, be pragmatic.

Questions

[Basic-1] – Pick a numeric attribute and compute the mean, median, standard deviation, min, and max values.
[Basic-2] – Pick a categorical attribute and determine the most/least common values.
[Basic-3] – Does your dataset suggest if more employees have left the company or remained in its employment?
[Basic-4] – Visualise an attribute with at least 3 distinct values. Briefly explain what your plot shows.

[Intermediate-1a] – Propose a question concerning the Attrition attribute and at least one numeric attribute, and answer it.

For example: Do people who leave the company, have on average higher rates or incomes? Briefly note if the answer is what you expected.

[Intermediate-1b] – Propose a question concerning the average Age and at least one categorical attribute, and answer it.
[Intermediate-2a] – Visualise Age against an interesting categorical attribute (not Attrition) and interpret the results.
[Intermediate-2b] – Visualise Age against an interesting numeric attribute (not Attrition) and interpret the results.
[Intermediate-3] – Does (and if so how does) Age relate to Attrition?

[Advanced-1a/b] – Visualise and correctly interpret the relationship of at least one categorical **and** one numeric attribute on Attrition. You may do this question twice, but **cannot** reuse attributes.
[Advanced-2] – Which attribute(s) appear to be most influential on Attrition? Be careful with missing values!!
[Advanced-3] – Convert Age into categorical that captures different types of employees (explain your choice of Age ranges) and identify: which type of employee seems to travel more.

[Show Off-1] – Build a logistic regression to predict Attrition and comment on its performance.
[Show Off-2] – Build a machine learning model, explain what it does and if applicable how it performs. Best (easiest) options here would be a C5.0 tree, Random Forest, Naïve Bayes, kNN, or a support vector machine. Be careful with missing values!!
[Show Off-3] – Run a clustering algorithm either on all numeric (e.g. k-means) **OR** all categorical (e.g. k-medoids) attributes.

Data Description

# The key to success in any organization is attracting and retaining top talent.
# You are an HR analyst at my company, and one of my tasks is to determine which factors
# keep employees at my company and which prompt others to leave. We need to know what
# factors we can change to prevent the loss of good people.

# You have data about past and current employees in a spreadsheet. It has various data
# points on our employees, but we're most interested in whether they're still with the
# company or whether they've gone to work somewhere else. And we want to understand how
# this relates to workforce attrition.

#Attributes:
# Age: in years
# Attrition: Y/N the dependent variable -- have they left the company?
# BusinessTravel: Non-Travel; Traval_Frequently, Travel_Rarely
# DailyRate: Consultancy Charge per Day
# Department: Human Resources; Research & Development; Sales
# DistanceFromHome: How far the employee lives from work
# Education: 1 'Below College'; 2 'College'; 3 'Bachelor'; 4 'Master'; 5 'Doctor'
# EducationField: Human Resources; Life Sciences; Marketing; Medical; Other; Technical Degree
# EmployeeCount: No of employees in this record
# EmployeeNumber: Employee ID
# EnvironmentSatisfaction: 4 point Likert scale: 1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
# Gender: Male / Female
# HourlyRate: Consultancy Charge per Hour
# JobInvolvement: 4 point Likert scale: 1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
# JobLevel Metadata not available -- make an assumption ☺
# JobRole: Healthcare Representative; Human Resources; Laboratory Technician; Manager; Manufacturing Director; Research Director; Research Scientist; Sales Executive; Sales Representative
# JobSatisfaction: 4 point Likert scale: 1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
# MaritalStatus: Divorced; Married; Single
# MonthlyIncome: monthly salary
# MonthlyRate: Consultancy Charge per Day
# NumCompaniesWorked: No. of previous employers
# Over18: Y/N
# OverTime: Yes/No
# PercentSalaryHike: Last Year's Increment
# PerformanceRating: 4 point Likert scale: 1 'Low'; 2 'Good'; 3 'Excellent'; 4 'Outstanding'
# RelationshipSatisfaction: 4 point Likert scale: 1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
# StandardHours: Contract hours
# StockOptionLevel: No available metadata -- make an assumption ☺
# TotalWorkingYears: Career Age
# TrainingTimesLastYear: No. of training courses attended last year
# WorkLifeBalance: 4 Point Likert Scale: 1 'Bad'; 2 'Good'; 3 'Better'; 4 'Best'
# YearsAtCompany: Time spent with company
# YearsInCurrentRole: Time in current role
# YearsSinceLastPromotion: No. of years since last promoted
# YearsWithCurrManager: Years spent with current manager

Expected Value of Questions

Question	Value	Notes
F1	up to 3	Based on level of correctness
F2	1	
F3	1, 2, or 3	Based on level attempted
F4	up to 3	There are many different ways to do this, some are better than others
Basic	1.1	
Intermediate	2.4	
Advanced	4	
Show-Off	6	

Possible Combinations of Questions in reference to potential grades

	Foundations				4 Questions					
	F1	F2	F3	F4	B	I	A	SO	Total (of 25)	%
Around Pass	3	1	1		4				9.4	37.6
	3	1	1		3	1			10.7	42.8
H2-2	3	1	2		2	2			13	52
	3	1	2		1	3			14.3	57.2
H2-1	3	1	2			4			15.6	62.4
	3	1	2			3	1		17.2	68.8
H1-1	3	1	3			2	2		19.8	79.2
	3	1	3			1	3		21.4	85.6
Exceptional	3	1	3				4		23	92
Max Score	3	1	3	1			1	3	32	128