Credit Scoring Assignment 2

S1889112

Ravi Patel

Question 1: (/4)

The Kolmogorov-Smirnov statistic is the greatest difference between two cumulative distribution functions, and as such the calculation is the same in development and monitoring. In scorecard development this difference is between the cumulative proportion of bads and goods. Hence in development we want a large KS statistic since we want to accept many goods and few bads. However in scorecard monitoring, the difference is between the expected proportion in each band (given a scorecard), and what is actually observed. An accurate scorecard means that there is a small difference between:

a) The proportion expected in each band
b) The proportion observed in each band

As such we want a small KS statistic. Therefore although the calculation is the same (rank observations by score, calculate the cumulative function, take the maximum difference), the interpretation is different (we want a large value for development, and a small value for monitoring).

Question 2: (/5)

A major disadvantage is that King's Buildings Bank cannot use product-specific information, as the scorecard must apply to both products. For example with a credit card business the bank can track a customer's transactions, the type of transaction, and any regular subscriptions that could take a toll on finances. Higher spending on credit cards is likely associated with a higher level of risk for the bank (i.e. a lower score), but this information cannot be included in the scorecard as it will be missing data for those who have the personal loan product. As such, only having one scorecard would mean increased risk for the bank (at least on the credit card product) due to the omitted data.

An advantage of a single scorecard is an increased amount of data for a given scorecard. Suppose the bank has 10000 datapoints, 7000 with the credit and 3000 with the personal loan. If 2 scorecards were built, the models would have 7000 and 3000 observations for the respective products. However if only a single scorecard is built, there are 10000 observations available, which will reduce the volatility of the model. This could be particularly useful for a "small local lender", as they likely do not have many observations after splitting on the product line. As such, the increased bias in model development is perhaps a worthwhile trade-off for the reduced volatility since the company can plan its future better.

Related to the issue of time is the issue of cost. Recalling that the bank is a "small local lender", it may be too high a risk to take on more employees, who will be needed to develop each scorecard, maintain them, and monitor them. In fact given the lack of data (as mentioned above), these employees may even be unnecessary, as separate scorecards may not provide much benefit due to the high model volatility from the lack of data. As such the employees required for this extra scorecard will have very little marginal benefit, and the marginal cost may even exceed the benefit they provide.

Question 3 (/5)

Overall, scores have increased over the 12 months, which has largely been a consistent trend at each percentile. There were early dips in the early stages of the 25th, 50th, 75th, and 95th percentiles. Conversely from November to December there has been an uptick in scores for all percentiles except the 50th and 75th (which has increased linearly since March). So in general there has been an upward shift in the score at each percentile. For the 5th and 10th percentiles this improvement was mostly between March and September, for the 25th between July and October, for the 50th and 75th consistently since March, for the 90th mostly in the final month, and for the 95th mostly in the final 2 months.

One reason for this increase in the scores could be a general economic improvement in Ireland. Assuming such a thing happened, on average people would have more income, and be more able to meet their financial obligations, meaning a lower probability of being a ‘bad’ loan, and hence a higher score.

Since scores are increasing, more applicants are above the cut-off, and as such more applicants are being accepted. A possible negative implication for the bank could come if the scorecard is underestimating the risk. This could expose the bank to large losses as it is now accepting more bad risks. As such the bank may want to validate the scorecard.

Question 4 (/12)

To apply the scorecard, I first copied the provided table into cells AW3:AX10. Then I inserted a column in AN (between "Model Ave # Transactions" and "Total Score"), and for the first row used "=VLOOKUP(M2, $AW$3:$AX$10, 2, 0)", which I dragged down to the bottom. To ensure the formula had worked, I used "=ISNA(AN2)" (dragged down), and summed the column to show there were no NA values (implying the lookup found all the values). I kept the old score as a means of comparison to the new score. The results of this procedure are shown in Table 4.1.

Table 4.1: Results from Including Unauthorised Debit Interest (L6M) Characteristic

Model Term	Model Ave CR T/O	Model Customer 1 Emp Status	Model Ave # Transactions	Model Months of Unauthorised Debit Interest (L6M)	Total Score (Old)	Total Score (New)	Status - G/B/R
27	43	43	29	4	142	146	R
27	50	58	29	4	164	168	R
18	32	43	29	-26	122	96	R
51	50	43	36	-9	180	171	R
18	16	43	17	-9	94	85	R
18	43	43	36	-9	140	131	R
18	50	43	36	-14	147	133	R
18	0	43	17	-37	78	41	R
18	0	43	17	-37	78	41	R

To set up calculations for the Kolmogorov-Smirnov (KS) statistic and the Gini coefficient, all we require is the new score, and the status – i.e. the final 2 columns. We take these columns into a new spreadsheet, and sort by the status to allow easy removal of the rejects. Once we have only accepted cases, we sort by the new score (ascending). We then calculate the cumulative good and bad proportions.

For the KS statistic, we want to get the maximum difference in the cumulative functions. However we also ensure that we take the maximum at a score where all cases have been accounted for. The reason we wish to do so is that in reality we cannot separate cases which have the same score. Without accounting for all the cases at a score, there is bias induced in the statistic by the ordering of goods and bads. For example, if by coincidence the bads all came first within a score, the KS would be higher due to random chance, and not any systematic feature of the data.

The first rows in calculating the KS statistic are shown in Table 4.2. The "%B-%G" column just takes the difference between the cumulative functions, whereas "All in Score" ensures that all cases with a given score have been accounted for. For example, the first 2 rows of "All in Score" are 0, as not all cases with a score of -37 have been accounted for. To get the KS statistic we take the maximum value of this column, which is 44.92%, corresponding to a score of 118. Note that the statistic before including this characteristic was 39.56%.

Figure 4.1: KS Graph

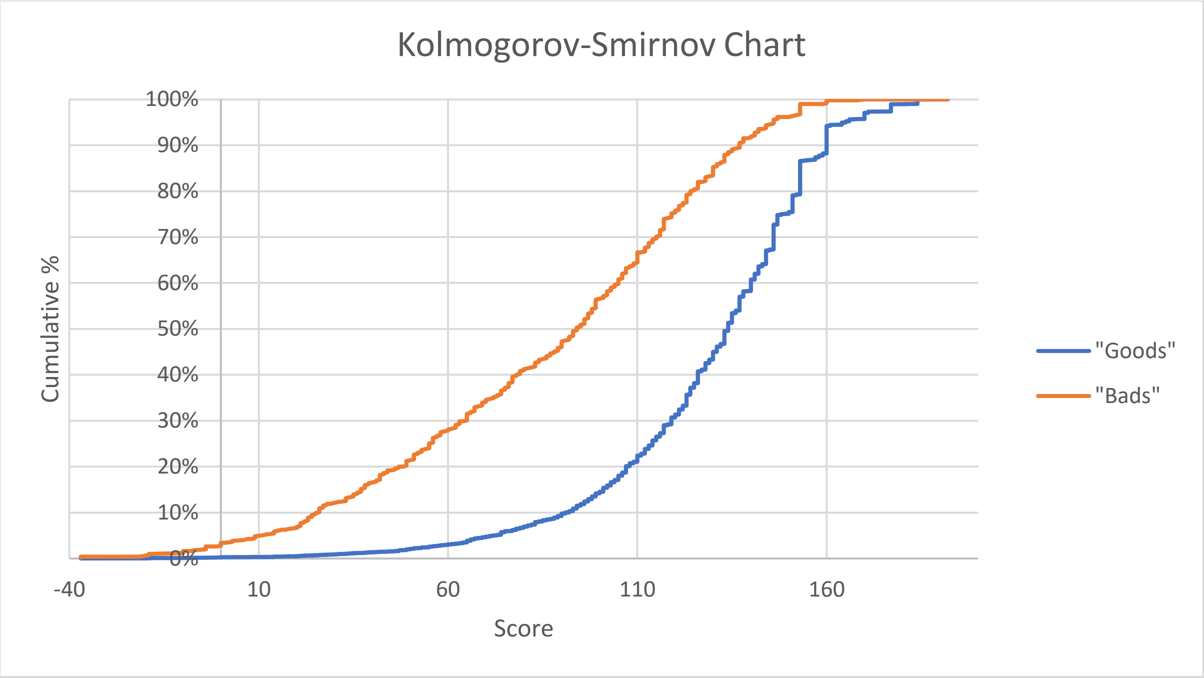
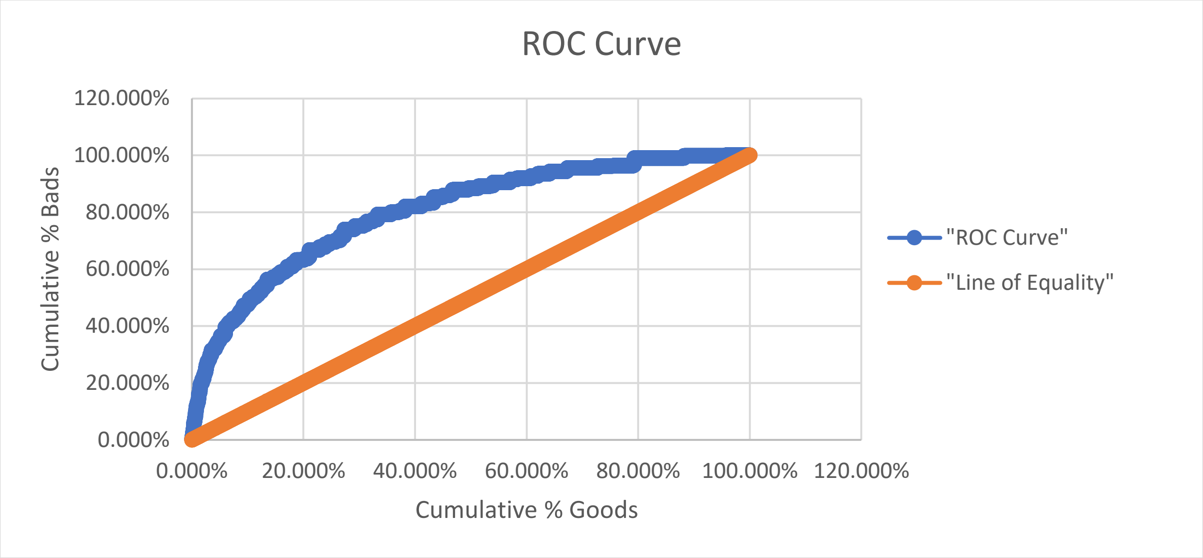


Table 4.2: KS Calculation

	Status - G/B/R	Cumulative Counts		Cumulative %		**KS**	
Total Score	(G/B/R)	Goods	Bads	%Goods	%Bads	%B-%G	All in Score
-37	B	0	1	0.00%	0.20%	0.20%	0.00%
-37	B	0	2	0.00%	0.40%	0.40%	0.00%
-37	G	1	2	0.02%	0.40%	0.39%	0.39%
-22	G	2	2	0.03%	0.40%	0.37%	0.37%
-20	B	2	3	0.03%	0.61%	0.57%	0.00%
-20	G	3	3	0.05%	0.61%	0.56%	0.56%
-19	B	3	4	0.05%	0.81%	0.76%	0.00%

To calculate the Gini coefficient (G), we use the AUC (area under the ROC curve). The equation to get the Gini coefficient is G = 2*AUC – 1. The ROC curve can be seen in Figure 4.2. The reflection of this along the line of equality is what we are more used to seeing in the Gini calculation. To get the area between the ROC curve and the line of equality, we calculate the area under the ROC curve and subtract 0.5 (the area under the line of equality). To normalise this area (as is required in a Gini calculation), we divide the area between the ROC curve and the line of equality by 0.5. This is where the result G = 2*AUC -1 comes from. The ROC curve comes from plotting “Cumulative % Bads” on the Y-axis against “Cumulative % Goods” on the X-Axis.

Figure 4.2 ROC Curve



The AUC can be calculated by calculating the area of each rectangle which forms the ROC curve (since bads and goods cannot increase at the same time, there are no diagonal points on the ROC curve so we do not require the area of a trapezium). This is done by multiplying the change in "Cumulative % Goods" by the level of "Cumulative % Bads". We sum all these areas to get the AUC. The first rows of the calculation can be seen in Table 4.3.

Table 4.3: Calculation of the Gini Coefficients

	Status - G/B/R	Cumulative Counts		Cumulative %		
Total Score	(G/B/R)	Goods	Bads	%Goods	%Bads	Area
-37	B	0	1	0.000%	0.202%	0
-37	B	0	2	0.000%	0.404%	0
-37	G	1	2	0.017%	0.404%	6.76E-07
-22	G	2	2	0.033%	0.404%	6.76E-07
-20	B	2	3	0.033%	0.606%	0
-20	G	3	3	0.050%	0.606%	1.01E-06
-19	B	3	4	0.050%	0.808%	0

For example in the 3rd row, the area is (0.017% - 0%)*(0.404%). Summing the area column yields a value of 0.80299. Hence the Gini coefficient is 2*0.80299 – 1 = 60.60%. Note that the Gini coefficient before including this characteristic was approximately 52.08% using the same method. The Gini coefficient can also be found using the Mann-Whitney U-Statistic which returns the same result, because of its own relation to the AUC (AUC = U/n_1n_2, where n_i is the sample size of group *i*. Here the groups are 'goods' and 'bads').

The new and old results for both the KS statistic and the Gini Coefficient are as follows:

Table 4.4: Pre and Post Comparison of KS and Gini

	Old	**New**	**Difference**
KS	**39.56%**	**44.92%**	**5.36%**
Gini	**52.08%**	**60.60%**	**8.52%**

The interpretation of both statistics is that a higher value indicates we can discriminate between good and bad cases better. As a result, we see that adding the characteristic of "Months of Unauthorised Debit Interest (L6M)" increases the KS by 5.4%, and the Gini by 8.5%. This is quite a large increase in both statistics, so I would recommend using this characteristic in the model due to the indication of increased ability to differentiate.

The problem of bias arises due to the fact we have not modelled on a random population. When scoring applicants, we want to determine the probability they are bad. However we have modelled the probability a customer is bad, **given** that they have already been accepted. I.e. we have modelled on a non-random sample. We have made the implicit assumption that all the rejects are bad, when in reality some would have performed well – in fact better than some that were accepted. In principle, we wish to 'swap' the bads we accepted for the goods we rejected.To do this, we need to see how the rejected population **would have** performed had they been given the chance. This is the idea behind reject inference, which is what we do to counteract the bias.

One way of doing this is through sample selection models. This involves modelling the missingness mechanism, which in this case is MNAR (missing not at random), as the chance person being rejected (i.e. unobserved) is correlated with the risk, since this is the basis we rejected them on.

The most common method is augmentation. We first build a model as normal (i.e. based only on accepts). Then we assume the accepted and rejected populations are identically distributed. That is the proportion of goods and bads falling into each score band are identical for accepts and rejects. We then assign 'good' and 'bad' labels to each member of the rejected population based on these proportions. Then we add the rejects (labelled as good or bad) back into the population, and rebuild the model.

Question 5: (/8)

Definitions

Applicant population: All prisoners who apply for temporary release as specified.

Reject: Prisoners who applied and were not granted release

Accept: The complement of the rejected population, split into goods and bads.

1) **Goods:** Those who were granted temporary release and were not involved in any incidents such as trying to escape, or committing any other crime
2) **Bads:** Those who were granted temporary release and **were** involved in any incidents such as trying to escape, or committing any other crime.

There could also be further constraints on accepted applicants to define their 'good' or 'bad' status. For example, associating with previous criminal contacts may not be against the law in and of itself, but it could be a requirement for release.

Characteristics

Firstly, consider the **time to release**. As an extreme example, suppose a convict was due to be released in 2 weeks. If they were granted temporary release over the weekend, the benefit of escaping is 2 weeks extra freedom, while the cost of trying to escape and being caught is likely a much longer stint in jail. As such the motive to try and escape is low. This alone means the time to release is positively correlated with the probability of being a bad (a longer time to release means a higher probability of trying to escape).

Next, consider the **reason for requesting release.** If a prisoner has the funeral of a close family member to attend, it is likely they are not just using this event as an excuse to try and escape, or commit another crime as they will be grieving.

A very important thing to consider is the chance that an offender is a danger to people once released. As such, we need to consider the **reason for incarceration.** Someone who was previously involved with gang activity may have some 'scores to settle', be in danger from other gangs themselves, and is generally more likely to be involved with violent crime. However someone who is in jail for financial crimes is likely not a material danger to the population. Additionally, someone who is in jail for a crime of passion assault is between these 2 risks.

Looking at the chance of reoffending when released, we look at the **number of criminal offences.** This characteristic would look at the number of criminal offences the person has committed in their lifetime. All else being equal, someone who has committed a crime once is less likely to reoffend than someone who has a history of reoffending.

We could also look at a **prisoner's behaviour** while incarcerated. Clearly, a prisoner who behaves well is more likely to be a good risk than someone who is consistently involved in fights, as it shows evidence of actual rehabilitation. For example this could be displayed just by not getting in trouble, but could be further enhanced by taking relevant courses, such as an anger management course for someone in jail for assault. The behaviour could be scored numerically from 1-10 based on an assessment of negative incidents (e.g. fights) any positive incidents (e.g. attending courses). Or if this is too much of a resource cost (getting an individual to assess the score based on the behaviour), subtract the number of negative incidents from the number of positive incidents. In this instance, a value of 0 would be neutral, a negative value would be bad, and a positive value would be good. This is a crude measure but would perhaps be a good starting point for a first scorecard.

Another possible characteristic could be the **score from a psychological evaluation.** As part of the application process, prisoners could be required to speak to a psychologist as part of an assessment, or fill out a form that attempts to assess their mental state. Though some psychological classifications (e.g. sociopathy, psychopathy) may not be 'orderable' in themselves, a psychologist may be able to define an ordering based on scores from forms or their own diagnosis.

A common characteristic used in applications of credit scoring (e.g. for banks, credit card organisations and in insurance) is an applicant's **marital status.** In the context of prison release, having a family on the outside could mean the applicant is less likely to risk an extended sentence, (or perhaps reduction in visiting rights if legal) by attempting to escape or commit another crime. An alternative characteristic that works along similar lines is the **number of children.**

Another universally used characteristic applicable in this context is **age.** Younger people may be more rash and impulsive, leading to a higher chance of trying to escape or getting angry and committing a violent crime. Additionally, younger prisoners may be more impressionable, and may be manipulated into committing a crime by peer pressure from previous criminal associates.

Question 6 (/4)

The idea behind stationary is that the transition probabilities between the states are constant over time. As such we look for reasons people may miss payments more or less often at certain times of the year, or for more or less people to request a loan (i.e. different numbers of new customers).

One reason people may default more is that they have higher expenditure at certain times of year (for example, many families go on holiday over summer or Christmas, and there is an increase in spending on presents over Christmas). Higher expenditure means people are less able to meet financial obligations, and are therefore more likely to miss payments which would mean more transitions to states such as 'Closed Bad', or to a higher number of delinquent payments.

A second reason is the flipside of this, which is that people may have more income at certain times of year (e.g. holiday pay over Christmas, especially for retail staff). As such people are more able to meet financial obligations, and are less likely to transition to states such as 'Closed Bad' or higher delinquencies.

Another factor that may cause variability over time is economic growth. Stationarity assumes the transition matrix is constant over time, which could mean the economic state of the world is constant. However economic growth has an upwards trend, and as such over the years we may get more people wanting credit cards, and being more able to pay them off on average. As such the probability of transitioning to 'Closed Bad' may reduce over time. In addition, economic volatility will likely have an adverse effect on the number of bad accounts (i.e. an increase in bad accounts), and fewer people will want credit cards, meaning a potential inability to replace those who go into absorbing states.

Finally, credit cards have no set end date, and account closures may be due to customers transferring to other lenders. As a result, the level of competition may play an important role. For example if there were only 2 lenders and one suddenly reduced their repayment rates, customers would (all else being equal, and assuming customers noticed this change) close their accounts with the more expensive lender and open accounts with the cheaper lender. As a result, for stationarity to hold there should be an expectation of no rate changes across lenders which is unlikely due to the constant competition and price reviews companies go through. Additionally, there may be strategic considerations such as one lender wanting to gear their book towards a certain attribute (e.g. build a book based on older people), and thus marketing directed at this group.

Question 7 (/5)

The idea behind this sort of analysis is checking the quality of the scorecard. The two curves should be aligned regardless of whether the characteristic is or is not in the scorecard, and as such we cannot tell if this characteristic is in the scorecard or not. This sort of analysis can tell us something about the characteristic whether it is or is not in the scorecard. If it is not in the scorecard, we may want to add it to the scorecard as it shows an ability to differentiate between goods and bads. If it is in the scorecard, we need to think of a transformation to apply such that the curves are aligned.

The reason these curves should be aligned is that the score is considered a 'sufficient statistic' for risk – this is the idea behind the score. That is, every piece of relevant information in determining in the risk is contained in the score, so individuals with the same score should pose the same level of risk. Even if every single characteristic is different, we expect 2 individuals with the same score to pose the same level of risk. As such, there should be no difference between individuals with or without a previous loan, provided they have the same score, as the same score indicates the same level of risk, and hence the same bad rate when averaging across the sample.

A drawback to adding 10 points is that this is not what the data itself is telling us, which may lead to misclassifying applicants in future. By adding 10 points to those with a previous loan, we underrate the risk of the attribute according to the data, since we are artificially inflating the score.

Question 8 (/2)

The purpose of a performance window is to see how applicants in a defined application window perform over time. Using the application window, we see how delinquency evolves over time, as more and more people become bad until a point this levels off. For payday loans there is not 'performance over time' element, as borrowers are expected to pay the loan back in one payment. As such, when given an application window, there is no development of delinquency. Rather each borrower's next payment determines whether they were good or bad.

Question 9 (/7)

To perform this calculation, we take a marginal χ^2 approach. To obtain the changes to be made to each score for actual and expected figures to align, we calculate the delta score (Δ) for each attribute, defined for attribute *i* as:

Equation 9.1: Δ Score

$$\Delta_i = (WoE_i - EWoE_i) * PDO/LN(2)$$

Where WoE is the weight of evidence given the actual data, EWoE is the weight of evidence given the expected data, and PDO is the points to double the odds that were used in the development of the scorecard (20 in this case). The weight of evidence attribute *i* is defined as:

Equation *9.2: Weight of Evidence*

$$WoE_i = LN\left(\frac{Goods_i}{Bads_i} * \frac{Bads}{Goods}\right)$$

The equation is analogous for EWoE. Using these equations, we can extend the given table to yield Table 9.1.

Table 9.1: Δ Score and Weights of Evidence

Loan term (Months)	Score	Actual Goods	Actual Bads	Expected Goods	Expected Bads	WoE (Actual)	WoE (Expected)	Δ Score
3-12	51	936	40	942.6	33.4	0.73	0.89	-4.78
13-48	27	4093	323	4088.6	327.4	0.11	0.08	1.05
49-60	18	1926	234	1952.6	207.4	-0.32	-0.21	-3.25
61-180	0	396	52	380.1	67.9	-0.40	-0.73	9.51
Totals		**7351**	**649**	**7363.9**	**636.1**			

The Δ score is the adjustment that should be made to the current score for the actual and expected results to align. We wish to keep scores as discrete numbers, so we round after applying the correction. The new scores are shown in Table 9.2.

Table 9.2: Proposed Scores

Loan term (Months)	Old Score	New Score
3-12	51	46
13-48	27	28
49-60	18	15
61-180	0	10

We see that the scores fall for “3-12” and 49-60”, rise by 1 for “13-48” (which is largely insignificant), and rise by 10 for “61-180”. This suggests that previously, those in “3-12” and “49-60” were not deemed risky enough, while those in “61-180” were deemed too risky when compared to reality. To check if the differences are statistically significant, we calculate the χ^2 statistic as follows:

Equation 9.3: Chi-Squared Statistic

$$\chi^2 = \sum_i 2 * \left(Goods_i * LN\left(\frac{Goods_i}{E(Goods_i)}\right) + Bads_i * LN\left(\frac{Bads_i}{E(Bads_i)}\right)\right)$$

We can then compare this to a chi-squared distribution with 3 degrees of freedom to determine whether the changes between the actual and expected values are significant. The calculations are shown in Table 9.3, with the above statistic highlighted.

Table 9.3: Chi-Squared Calculation

Loan term (Months)	**Score**	**Actual Goods**	**Actual Bads**	**Expected Goods**	**Expected Bads**	**Marginal χ^2**
3-12	51	936	40	942.6	33.4	1.27
13-48	27	4093	323	4088.6	327.4	0.06
49-60	18	1926	234	1952.6	207.4	3.64
61-180	0	396	52	380.1	67.9	4.71
Total		7351	649	7363.9	636.1	**9.68**

Our test-statistic is therefore 9.68, which can be tested against a χ_3^2 using the Excel function “CHISQ.DIST.RT” to get the probability of getting this particular test-statistic or higher. That is, we use “=CHISQ.DIST.RT(9.68, 3)”, which yields a p-value of 0.0214, which suggests the changes are statistically significant at the 5% level. Consequently, we should make the proposed changes as per Table 9.3.

Question 10 (/13)

Question	11A	11B	11C	11D	11E	11F	11G	11H	11I	11J	11K	11L	11M
Answer	B	B	D	C	B	A	D	D	B	D	E	E	B