MTH 325, Final Exam, Spring 2024 Name ____

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:
Use the work you did at home to answer these questions about spending.

1. Based on your correlation table or correlation plot, identify the variable that has the highest negative correlation with Amount Spent. What is the (approximate) correlation value?

2. Based on your ANOVA model, should interactions be included in your model or not?

3. Do the residuals from your ANOVA or general linear model appear to be normally distributed?

4. After converting the categorical variables to dummy variables, which two variables appear to have the highest correlation (positive or negative)?

5. After performing backward selection, what is the R^2 value of your resulting model?

6. Write the equation you obtained from your backward selection process for predicting operating expenses. Be sure to clearly indicate what each variable in the equation represents.

7. Describe how your other model selection methods differed (or were similar to) the results obtained from the backward selection process.

8. What percentage of the variability in Amount Spent can be explained by the relationship to the other model variables?

9. Answer this question and the remaining questions in Part 1 using the backward selection model you found by hand. Do your diagnostic plots suggest any outliers or model problems? Explain.

10. How do your predictions for the 10 extra people? How does your residual error (RMSE) differ from the model residual error?

11. Interpret the meaning of the Married coefficient in the context of the problem.

12. If you needed to build a model of Amount Spent with two variables, what would they be, and why?

Use the work you did at home to answer these questions about the time series model.

13. Does the model appear approximately stationary or does there appear to be a trend? Consider any boxplots or histograms here, as well as any time series plots or decompositions you may have done.

14. Based on your PACF graph, how many lags should be included in your time series model? Why?

15. What settings did you use for your ARIMA model? Why? What diagnostics did you use to select these settings?

16. Write the equation of your final time series model.

17. What is the AIC of your final model? How good does the model appear to fit?

Part II:

18. Recall that $Cov(X, Y) = E(XY) - E(X)E(Y)$. For the probability density function $f(x, y) = \frac{1}{2}x^2(y + 1), y \in [0,1], x \in [0,1]$, find the covariance.

19. Consider the small data set $\{(3,2),(5,4),(9,9)\}$. Find the value of the regression coefficients for $y = \beta_0 + \beta_1 x$, using the normal equation $(A^T A)^{-1} A^T Y = B$. Write the coefficients you find in the equation.

20. Suppose that a classification model (such as logistic regression) produced the following confusion matrix. Calculate the accuracy and discuss whether the model result reveals any potential problems.

	Yes	**No**
Yes	641	11
No	24	2

21. Describe clustering (in machine learning) and give an example of a machine learning algorithm that implements this learning method. Is this method an example of supervised, unsupervised or semi-supervised learning.

22. Describe how Gaussian process regression works in general terms.

23. What are some reasons it might be beneficial to use a non-parametric nonlinear model for a regression problem rather than a parametric non-linear model?

24. What is one reason you might get an error from the decompose() function applied to a time series?

25. Explain why autocorrelation prevents us from using traditional regression to model some time series data.

26. Why are irregular time series so much more difficult to work with than regular time series? Describe some methods we can use to make irregular time series more regular.

27. How do we use the AUC (of an ROC curve) as a diagnostic for a classification model?

28. Describe the difference between LASSO and Ridge regression. Explain how the penalty helps to address the bias-variance trade-off.

29. How does Spearman’s correlation differ from Pearson correlation?

30. What are some potential advantages and disadvantages of using variable transformations in a model?

31. What is the difference between an outlier and an influential point in a regression model?