**Problem 1**

**[T2 HOTELLINGS]**

To verify that exist a difference between the two store we create at first a dataset with the difference of the mean price and conditions for the same product.

To perform a test we have to check that the data belong to **(multivariate)** normal distribution. We perform a Shapiro test and we obtain a p-value of **[big p-value]**, so we can assume normality.

We perform a T2 test with the following hypothesis:

H0 : mean = H1: mean !=

We obtain a p-value of **[small p-value]**, so we at level **[alpha]** reject the null hypothesis and we can say that exist a difference between the two stores.

**[Ellipse]**

We draw the elliptical confidence region at level **[level]** on the mean:

[PLOT ELLIPSE]

The parameter of the ellipse are: center in **[Coord**], direction of the principal axis **[direction]** and length **[Length]**.

The equation is the following:

Dove:

**[T2 simultaneous CI]**

We already checked the gaussianity of data so we can perform a simultaneous T2 test at level [**level**] on the **[Variables]** and the result is the following:

[Result]

So … [Conclusion]

**[PCA]**

We perform a Principal Component Analysis to explain the variability of the dataset. We will find the directions around with the maximum variability is spread.

[RESCALE] As we can see in the box plot of the data their scale is too different to be compared so we rescale them before studying.

The first [TREE] principal components explain respectively [VAR PERCENT] of the variance and their loadings are the following:

[PLOT LOADINGS]

As we can see [INTERPRETATION]

The screeplot show the distribution of the variance along the principal components. As we can see

**[Clustering]**

We look at our data and suppose that there are **[number]** clusters.

[PLOT OF DATA]

The we compute the distance matrix using **[DISTANCE]** distance and using a **[TYPE]** linkage we build the dendrogram. The cophenetic coefficient is: [COEFF] so we expect a NOT noisy dendogram

[DENDOGRAM and COLORED SCATTERPLOT]

As we can see the clusters are **[DESCRIPTION]**.

The center of the clusters are: [CENTER] and their dimension are: [DIMENSION].

**[ANOVA/MANOVA]**

We check the assumption for the ANOVA/MANOVA

1. Normality (multivariate) in each group:

We get the following p-values [BIG P\_VAL] from Shapiro test so the assumption are satisfied.

1. Same covariance structure in each group:

By visual inspection we check this assumption

We perform a Bartlett test, and we get a p-value of [BIG P-VALUE] so we can’t reject the null hypothesis of equality between the variance

We perform the anova test using as group the cluster label and the Wilks statistic suggests us that there’s a sensible difference between the mean of the groups. The p-value is nearly 0 (< 2.2\*10^(-16)) so there is difference between the mean of the groups.

**[Bonferroni]**

**(Mean1 – Mean2) - sqrt(diag(Sp)\*(1/n1+1/n2)) \* t(1 - alpha/(p), n1+n2-1-p)**

We perform [K] simultaneous test [ON THE MEAN] using Bonferroni correction to pursue a global level of [LEVEL], the result are the following.

[BONFERRONI INTERVALS]

As we can see [INTERPRETATION]

**[LDA and QDA]**

We want to classify [SUBJECT] so we decide to create a LDA model or QDA model. At first we start by verifying their assumptions, the assumption for LDA are the followings:

1) The data inside group are normal  
2) Equal misclassification costs  
3) The variance of the groups is the same

While for QDA the third is not necessary.

We perform a Shapiro test on the groups and the resulting p-values are: [BIG P-VAL] so the normality assumption is met.

The equal misclassification cost is the same in this problem but the variance of the groups is not the same as shown from the Bartlett test which return the p-value of [SMALL P]. This means that the hypothesis of equal variance is rejected.

We the decide to use a QDA model assuming a priori probability of [PRIOR]. The group means and varinces are the following

[MEAN AND VARIANCES]

And the plot of the classification region is:

[PLOT OF REGION]

We then compute the AER by leave one out CV and the result is: [SMALL AER] which seems good.

If we calculate the posterior probability based on the training set of the classifier are:

[TOTAL PROB THM]  
IF we want to predict the point d the result is: [CLASS] with this posterior probability: [POST PROB]

**[Linear regression]**

We are going to study [FENOMENON] using the designed linear model.

The plot of the data is the following: [PLOT]

We fit the model on our training set and obtain the following parameters:

[PARAMETERS remember SIGMA (not O^2)]

[PLOT OF THE MODEL RESIDUAL]

We can trust the model since by looking at the residual we see no patterns so the homoscedastic assumption is met; they seem to be gaussian and there are no leverage points. The Shapiro test on the residuals gives a p-values of [BIG PVAL] so we confirm the gaussianity intuition on the residuals.

The VIF values are: [SMALL VIF] so we can say that there is not much collinearity between variables.

LINEAR HYPOTESIS We perform statistical tests using Linear Hypothesis to see if [VARIABLE] is significant. The p-value of the test is [BIG P] so we accept the hypothesis of null coefficient at level [LEVEL] and we remove the variable from the model.

The new model has the following coefficient: [COEFFICIENT]

[PLOT OF MOD] and we can see that the model assumption are still met as before and we have an improvement of the R2 from [VAL] to [VAL].

We predict the value for the new point [POINT] using our model and the resulting confidence interval at level [LEVEL] is: [CI]

**[GEOSTAT]**

We are going to study the spatial [FENOMENON].

We check the assumption of isotropy and second order stationarity:

[VARIOGRAM OF DIRECTIONS]

The variogram along the difference dimension is similar and the convergence levels seem the same so we can assume isotropy and secondo order stationarity.

The resulting fitted variogram has the following parameter: [PARAMETER] and is displayed below:

[FITTED VARIOGRAM]

BLUE = TRUE PER stimare coeff o se vogliamo usare il modello senza il delta (dipendenza da dati spaziali)

[YES] The model assumption of second order stationarity and isotropy are satisfied since we can see that the variogram seem to converge to a horizontal asymptote. We can then use Ordinary kriging to make prediction.

[NO] The model assumption of isotropy is satisfied but the second order stationarity is not since we can see that the variogram seem diverge as the abscissa grows. We can then use Universal kriging to make prediction.

Fitting the designed model [MODEL] on our training set via generalized least square we obtain the following parameters: [PARAMETERS remember SIGMA (not O^2)].

We now predict the [VARIABLE] for a new point with coordinates [COORD]. Using [UNIVERSAL] Kriging we obtain the value of: [PREDICTION] and an estimate variance of [VAR]. We can’t trust this variance since in the hypothesis of [UNIVERSAL] Kriging Sigma is supposed to be known while here we estimated it, so the variance is smaller than the real one.

**[FUNCTIONAL smoothing]**

From the recorded data we want to find the number of bases which best interpolate them avoiding undersmoothing or oversmoothing. We use generalized cross validation to quantify the problem and find the best number of bases of type [TYPE] and degree [DEGREE]. In this case the best number is [NUMBER].

[PLOT of basis and PLOT OF SMOOTHED]

LS Smoothing can be seen as a linear regression where the design matrix results from the evaluation of the bases functions in the observed points. So, once we have the value thanks to the LS we can estimate the variance and compute confidence intervals for the choose function.

[PLOT Of confidence intervals]

To evaluate the first derivative from data we use an approximation derived from the centered finite difference or in the case of smoothed data we can evaluate the derivative of the smoothed function.

[PLOT OF different type of derivative]]

**[FUNCTIONAL PCS]**

We have a dataset of [DESCRIPTION].

[PLOT FUNCTIONS]

To smooth the data at first we create a set of [NUMBER] basis of the type [TYPE]. We project the orginal functions on them and obtain the following result:

[PLOT of SMOOTHED]

We can now perform Principal component analysis on the data to find the function around which the maximum variability is spread. The first [TREE] PCS explain respectively [VALUES] so summed the: [VAL].

Here we can see the screeplot resulting and the plot of the first [TREE] eigenfunctions.

[SCREEPLOT and PLOT OF HARMONICS]

We decide to take just the first [NUM] PCS because the explain the [PERC] of the variance.

As we can see from the previous plot the first [EXPLANATION …]