Université de Paris

Master 2 IMB
2020-2021

ETUDE DES FORFAITS DE SKI

Rafi Rashid – Ayour Arakchou
Université de Paris – *45 rue des saints-pères 75006*

SOMMAIRE

Partie 1 : Présentation et analyse univariée des données

1.1 Présentation des données

Nous avons à notre disposition une base de données de 208 observations caractérisées par 13 variables.

Les variables qualitatives

Au nombre de deux, elles nous donnent une information sur le nom de la station de ski (variable « NomStation ») et sur la localisation régionale de cette dernière (variable « Region »).

Les variables quantitatives

A l'inverse des variables qualitatives, ces variables peuvent faire l'objet d'opération dans le but d'obtenir des indicateurs statistiques comme la moyenne, la médiane, l'écart-type…

Voici un court descriptif de ces dernières :

Nom des variables	Descriptif
Prixforfait	Prix du forfait choisi (en euros)
Vertes	Nombre de pistes vertes (débutant)
Bleues	Nombre de pistes bleues (intermédiaire)
Rouges	Nombre de pistes rouges (difficile)
Noires	Nombre de pistes noires (expert)
NbPistes	Nombre de pistes de ski dans la station
AltitudeMin	Altitude minimale de la station
AltitudeMax	Altitude maximale de la station
Denivele	Nombre de dénivelé
Telesieges2places	Nombre de télésièges à 2 places
Telesieges4places	Nombre de télésièges à 4 places
Teleskis	Nombre de téléskis
Remontees	Nombre de remontées

1.2 Analyse univariée

Dans cette partie, nous allons présenter les résultats obtenus à l'issue de l'étude descriptive des variables en distinguant les variables quantitatives des variables qualitatives.

Comme nous pouvons le voir dans la table 1, nous disposons de 13 variables qualitatives pour 208 observations. Le prix moyen d'un forfait est environ égal à 150 € avec une valeur maximale pouvant atteindre 285 €. L'altitude minimale et maximale des stations est en moyenne de 1313 et 2150m respectivement.

Variable	N	Moyenne	Ecart-type	Minimum	Maximum
Prixforfait	208	149.8605769	61.5969820	0	285.0000000
Vertes	208	7.0384615	6.1407376	0	45.0000000
Bleues	208	13.8701923	14.8661492	0	71.0000000
Rouges	208	12.2115385	11.8486209	0	91.0000000
Noires	208	4.2163462	4.7717285	0	34.0000000
NbPistes	208	37.3365385	34.5888510	1.0000000	235.0000000
AltitudeMin	208	1313.87	293.7324050	600.0000000	2300.00
AltitudeMax	208	2150.44	583.3446425	970.0000000	3600.00
Denivele	208	840.0528846	494.2713977	60.0000000	2350.00
Telesieges2places	208	0.7451923	1.6875434	0	14.0000000
Telesieges4places	208	2.1923077	3.2662139	0	22.0000000
Teleskis	208	10.3750000	8.1075947	0	60.0000000
Remontees	208	18.2644231	17.9812646	0	108.0000000

Table 1. Indicateurs statistiques des variables quantitatives

La figure 1 nous montre que le troisième quartile est environ égal à 195 €, c'est-à-dire que 75% des forfaits de ski ont un prix inferieur à cette valeur tandis que le premier quartile avoisine les 100 €. De plus, le prix de forfait médian est égal à 150 € ce qui signifie qu'une station de ski sur deux propose un forfait supérieur à ce montant. A noter également qu'il n'y a pas de valeurs aberrantes.

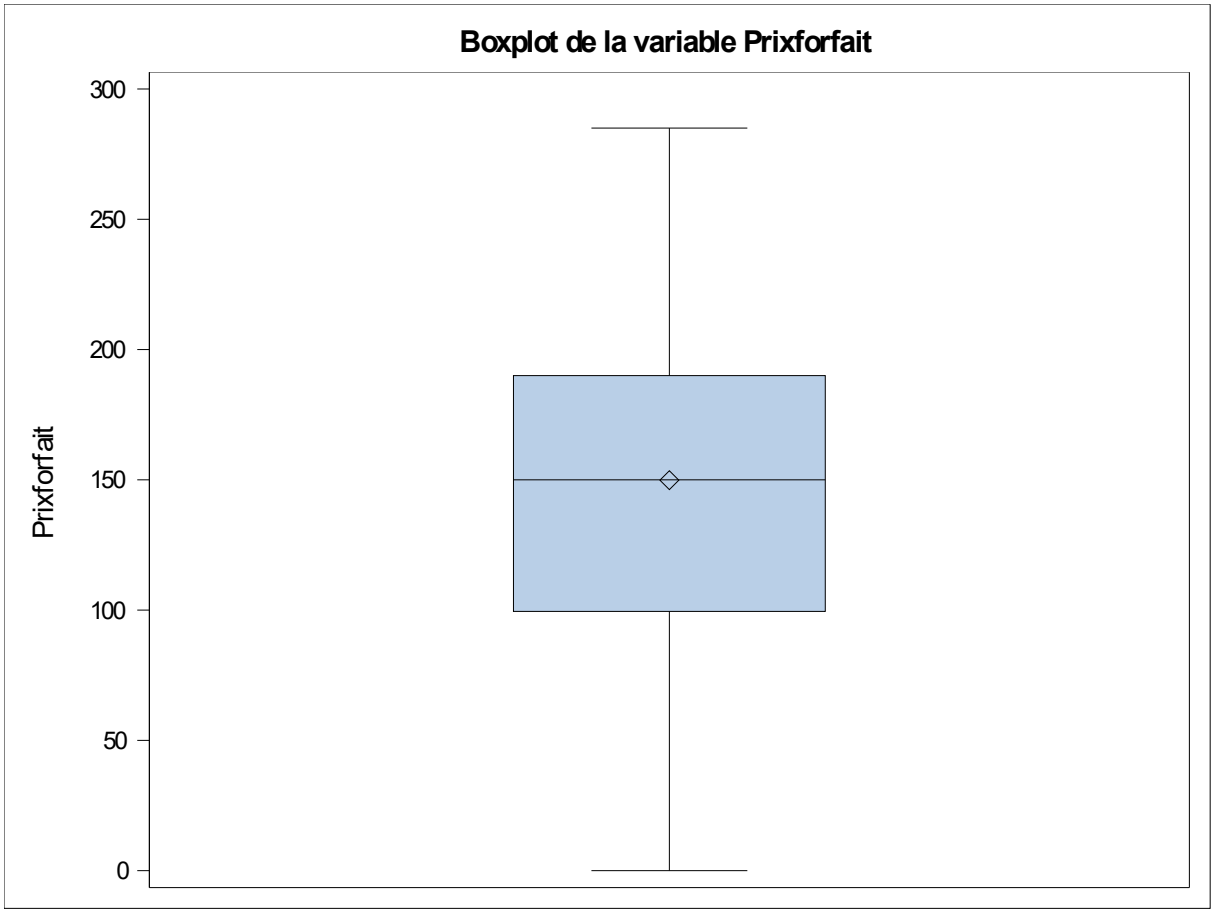


Figure 1. Boîte à moustaches représentant la distribution du prix du forfait

La variable Region est quant à elle une variable qualitative. On observe une part importante de stations de ski des alpes du nord (52.4%) probablement en raison du climat et de la géographie propice à la pratique du ski. A l'inverse la région du Jura représente que 3.37% de nos stations de ski.

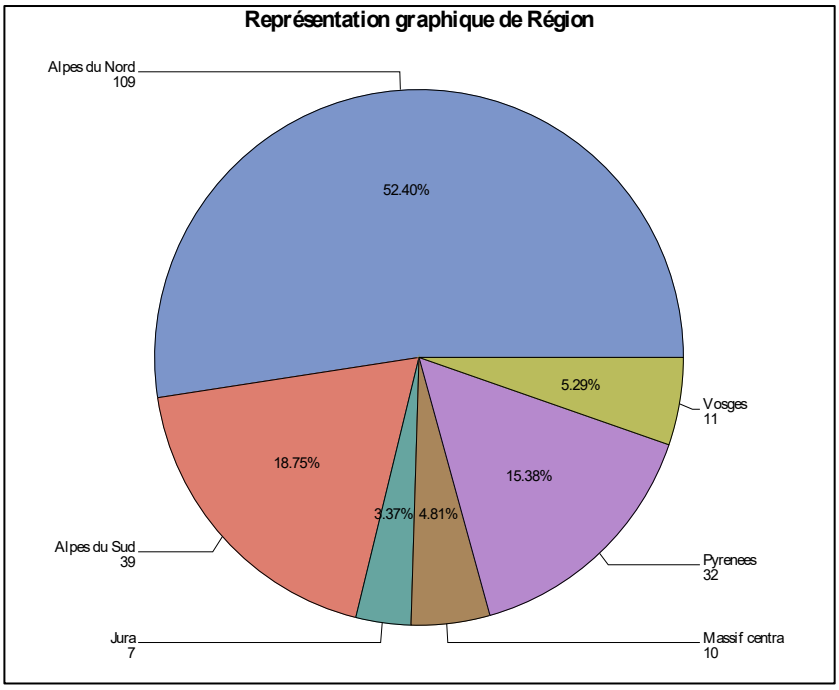
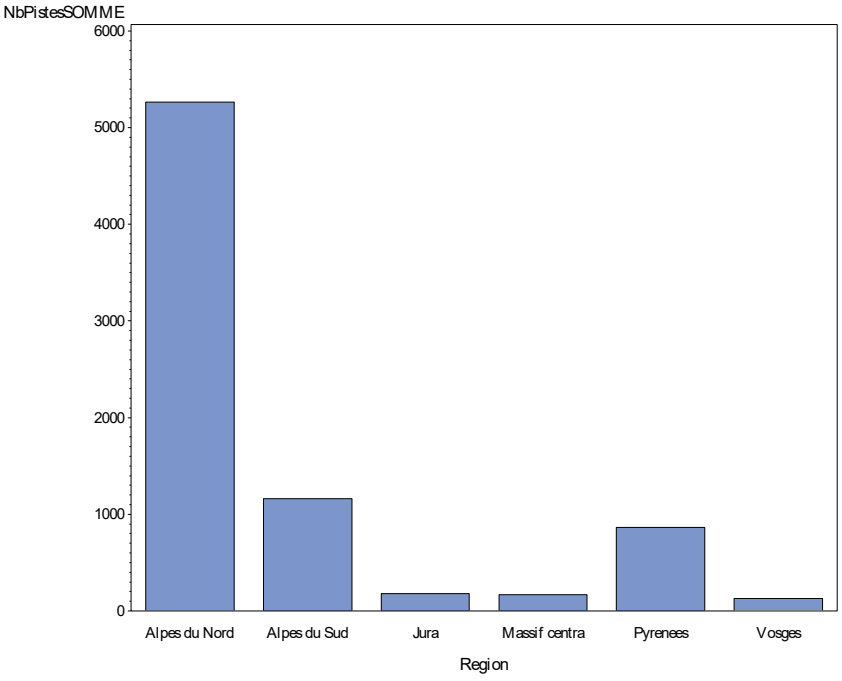


Figure 2. Pie chart de la variable Région



Il est intéressant de représenter le nombre de pistes des stations de notre jeu de données selon la région. Ainsi, on voit bien que les alpes du nord comptent énormément de pistes (toutes couleurs confondues) avec un total de près de 5200 pistes. Les alpes du sud qui représentait tout de même presque 20% des stations du jeu de données ne comptent par exemple qu'environ 1200 pistes.

Figure 3. Distribution du nombre de pistes selon la Region

Pour terminer le traitement univarié de nos variables, nous avons réalisé des tests d'ajustement afin de savoir quelles variables pouvaient être approchés par une loi gaussienne (dont les paramètres sont estimés à l'aide des informations à disposition). Aussi nombreux sont-ils, le logiciel SAS utilise 4 tests différents : le test de Shapiro, le test de Kolmogorov-Smirnov, le test de Cramer-von Mises et le test d'Anderson-Darling. Ces derniers sont basés sur la confrontation de l'hypothèse nulle, notée H0, et l'hypothèse alternative, notée H1. Soit :

H0 : La variable étudiée peut être ajusté par une loi gaussienne.

H1 : La variable ne peut pas être ajusté par une loi gaussienne.

	Statistique de test				P-valeur				
Variables	Shapiro-Wilk	Kolmogorov-Smirnov	Cramer-von Mises	Anderson-Darling	Shapiro-Wilk	Kolmogorov-Smirnov	Cramer-von Mises	Anderson-Darling	Normalité
Prixforfait	0.978371	0.064117	0.136455	1.092075	0.0027	0.0363	0.0377	0.0075	Rejetée
Vertes	0.781883	0.163162	1.640998	9.919946	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
Bleues	0.755793	0.203247	2.546227	14.77097	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
Rouges	0.789532	0.180395	1.911435	10.67377	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
Noires	0.768716	0.195966	2.178334	12.60609	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
NbPistes	0.800514	0.156736	1.855154	10.85675	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
AltitudeMin	0.986792	0.068185	0.151856	0.905007	0.0502	0.0191	0.0229	0.0217	Rejetée
AltitudeMax	0.98368	0.079068	0.153551	0.906056	0.0166	<0.0100	0.0220	0.0216	Rejetée
Denivele	0.945728	0.0827	0.414714	2.797087	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
Telesieges2pl	0.460559	0.329395	6.289584	32.28732	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
Telesieges4pl	0.666584	0.251044	3.157343	18.252	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
Teleskis	0.790819	0.172292	1.872322	10.54888	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée
Remontees	0.763181	0.184723	2.55492	14.24261	<0.0001	<0.0100	<0.0050	<0.0050	Rejetée

Figure 4. Test d'adéquation à la loi gaussienne.

Sur la table présente ci-dessus on retrouve l'ensemble des résultats obtenus par nos tests. On remarque que pour chaque variable, la conclusion est la même. En effet, si on fixe notre seuil α à 5%, on s'aperçoit que l'ensemble des p-valeurs sont strictement inférieures à ce dernier. On rejette donc l'hypothèse nulle. Ainsi, pour chaque test, la conclusion est la même : la variable étudiée ne peut être ajustée par une loi gaussienne.

Partie 2 : Etude statistique du prix du forfait de ski

2.1 Etudes bivariées

Au sein de cette première partie, nous allons recenser les variables qui expliquent le mieux le prix du forfait de ski. Cette analyse bivariée va s'appuyer sur la construction de modèle de régression linéaire simple, des tests de corrélation pour les variables quantitatives et sur des analyses de la variance (ANOVA) pour les variables qualitatives.

Les variables quantitatives :

L'objectif de cette section est de cibler les variables quantitatives qui ont un lien avec la variable d'intérêt, le prix du forfait.

Nous avons alors construit les modèles suivants :

• Modèle 1 – Le prix du forfait en fonction de son nombre de pistes. Dans le but de juger de la qualité de notre modèle, nous avons effectué trois différents tests statistiques :

• Le test de Pearson qui confronte les hypothèses suivantes :

- H0 : La corrélation est nulle.

- H1 : La corrélation est non nulle.

• Le test de Student qui confronte les hypothèses suivantes :

- H0 : La pente de régression est nulle.

- H1 : La pente de régression est non nulle.

• Le test de Fisher qui confronte les hypothèses suivantes :

- H0 : Le modèle yi = β +εi.

- H1 : Le modèle yi = αxi + β + εi.

Les variables qualitatives :

Notre jeu de données ne comporte que deux variables qualitatives : Nomstation et Region. Pour la suite, nous allons nous focaliser uniquement sur la région, le nom de la station n'apportant pas vraiment d'information pouvant influencer le prix du forfait de ski.

Étude du prix d'un forfait de ski en fonction de la région d'appartenance de la station de ski.

$$\forall (i,j) \quad y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad avec \quad \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

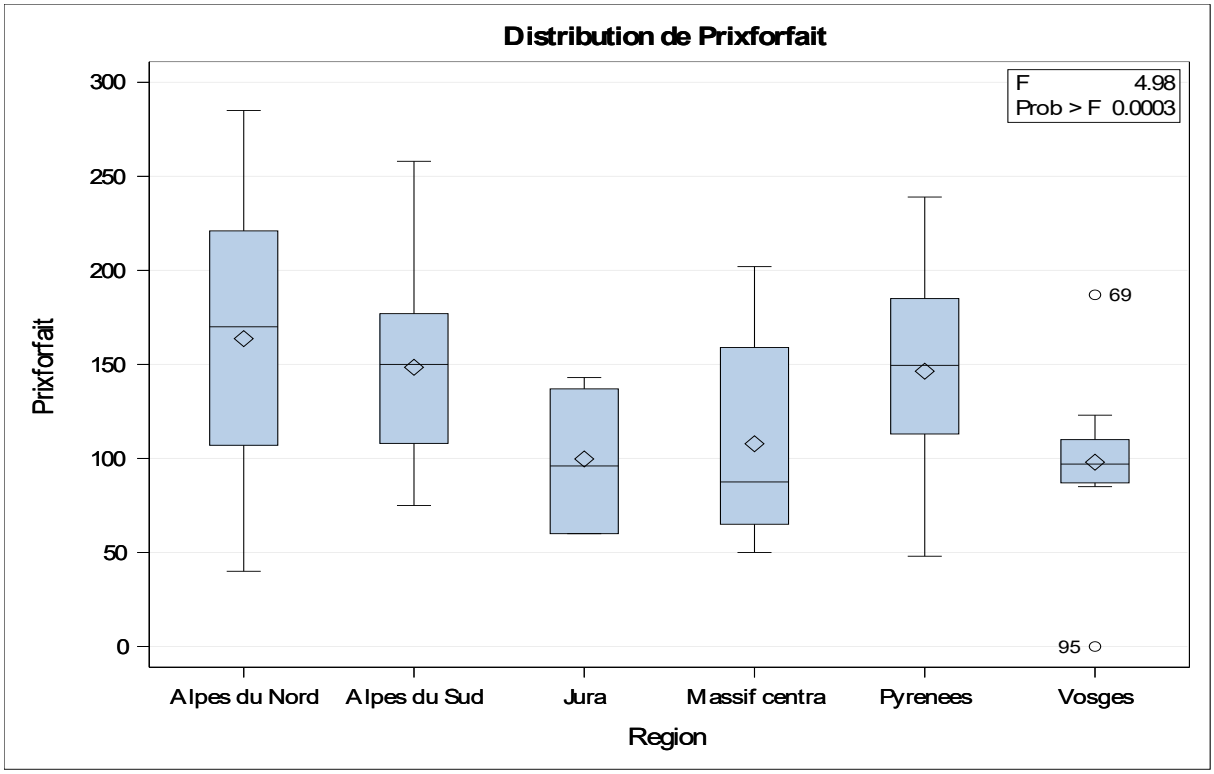


Figure 5. Boîtes à moustaches du prix du forfait selon la Region de la station de ski.

Premièrement, comme nous pouvons le constater, la distribution du prix du forfait selon la région est assez similaire pour les régions Alpes du Sud et Pyrénées, de même que pour Jura et Massif central. De plus, on remarque une médiane plus importante pour la région Alpes du Nord. Enfin on peut noter la présence de données atypiques pour la région Vosges. Au vu de ses observations, il semblerait bien qu'il existe un lien significatif entre la région et le prix du forfait de ski.

Comme pour les variables quantitatives, nous allons mettre en place un test statistique afin de savoir si le modèle construit explique de façon significative le prix du forfait de ski.

Pour ce faire, nous avons utilisé le test suivant :

• Le test d'ANOVA qui confronte les hypothèses suivantes :

- H0 : $\alpha 1 = \ldots = \alpha 6 = 0$

- H1 : $\exists i$ tel que $\alpha i \neq 0$

Après avoir dichotomisé notre variable Region via une étape DATA, nous appliquons une proc reg et obtenons une statistique de test F = 4.98 et une p-valeur = 0.0003 < 5%. La p-valeur étant strictement inférieure à notre seuil (fixé à 5%), on rejette l'hypothèse nulle. Ainsi, la région explique de façon significative le prix du forfait. De plus, nous avons un $R^2 = 0.10$, ce qui est très faible. On en déduit que la modélisation du prix du forfait en fonction de la région est de très mauvaise qualité. Nous ne chercherons pas à le valider par la suite

2.2 Etude sur le prix du forfait selon des catégories de variables

A présent nous allons étudier spécifiquement le prix du forfait en fonction d'un groupe de variables. Ces groupes de variable ont été construit sur les informations que porte chaque variable. Le but étant de rassembler les variables qui caractérisent le même aspect du forfait entre elles. Deux groupes ont été construits :

- Les pistes (vertes, bleues, rouges, noires). Nous n'incluons pas la variable NbPistes puisqu'elle résulte de la somme des variables vertes, bleues, rouges et noires.

- Les caractéristiques naturelles (Altitudemin, Altitudemax, Denivele).

- Les infrastructures de la station (Telesieges2places, Telesieges4places, Teleskis, Remontees).

Les aspects caractérisant les pistes accessibles impactent-ils vraiment le prix du forfait ?

L'étude du prix du forfait en fonction des variables caractérisant les pistes nous mène à la table de résultat suivante (utilisation de la proc reg) :

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	103.70431	4.42000	23.46	<.0001
Vertes	1	0.43380	0.68685	0.63	0.5284
Bleues	1	1.64246	0.38748	4.24	<.0001
Rouges	1	0.32757	0.59598	0.55	**0.5832**
Noires	1	3.87103	1.25690	3.08	0.0024

Table 2. Modèle complet

On décide de retirer la variable « Rouges » puisque le test de Fisher est non significatif. L'impact de cette variable sur notre prix du forfait est donc négligeable.

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	103.93187	4.39303	23.66	<.0001
Vertes	1	0.58060	0.63170	0.92	**0.3591**
Bleues	1	1.74296	0.34104	5.11	<.0001
Noires	1	4.19011	1.11288	3.77	0.0002

Table 3. Modèle sans « Rouges »

De nouveau, on retire cette fois la variable « Vertes ».

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	105.74235	3.92515	26.94	**<.0001**
Bleues	1	1.79996	0.33523	5.37	**<.0001**
Noires	1	4.54242	1.04439	4.35	**<.0001**

Table 4. Modèle final

Parti du modèle « Prixforfait = vertes + bleues + rouges + noires + un terme de bruit », certaines variables présentaient un test de Fisher non significatif ce qui voulait dire qu'elle pouvait être retirée du modèle. Ainsi, nous avons procédé par étape et avons retirés les variables une à une. L'objectif étant d'obtenir un modèle dans lequel toutes les variables ont un test de Fisher significatif (p-valeur < 5%).

Root MSE	40.88496	**R carré**	**0.5637**
Moyenne dépendante	149.86058	**R car. ajust.**	0.5594
Coeff Var	27.28200		

Table 5. R carré et autres indicateurs statistiques sur le modèle

Pour conclure, le modèle construit explique de façon significative le prix du forfait. De plus, le R^2 est égale à 0.5637 ce qui signifie que notre modèle explique environ 56% de la variance du prix du forfait.

Comment se comporte les variables caractérisant l'aspect naturel de la station selon le prix du forfait ?

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	9.95044	11.76920	0.85	0.3988
AltitudeMin	1	0.04333	0.00828	5.23	**<.0001**
Denivele	1	0.09877	0.00492	20.08	**<.0001**

Table 6. Modèle complet

Par le même procédé, nous sommes partis du modèle « Prixforfait = AltitudeMax + AltitudeMin + Denivele + un terme de bruit ». Seule la variable a été jugée négligeable ce qui explique qu'on garde uniquement les variables AltitudeMin et Denivele.

Root MSE	34.96742	**R carré**	**0.6809**
Moyenne dépendante	149.86058	**R car. ajust.**	0.6777
Coeff Var	23.33330		

Table 7. R carré et autres indicateurs statistiques sur le modèle

De plus, nous avons un $R^2 = 0.6809$ ce qui est bien meilleur que dans le modèle précédent.

Qu'en est-il des variables caractérisant les infrastructures de la station ?

De nouveau, on applique la même méthode et seule les variables Telesieges2places, Teleskis, et Remontees sont retenu dans ce modèle. Le R^2 s'élève à 0.5807 ce qui signifie que 58% de la variance de Prixforfait est expliquée par ce modèle.

2.3 Etude générale

Pour finaliser l'étude du prix des forfaits, nous avons créé un modèle de régression général relatant l'ensemble des variables mises à notre disposition expliquant de façon significative le prix des diamants de notre jeu de données. Ce dernier a été construit de telle sorte à avoir un compromis nombre de variable/qualité de modélisation. L'objectif étant de sélectionner judicieusement les variables du modèle afin d'avoir le moins de variables explicatives possible sans affecter la qualité de modélisation. Pour ce faire, on a utilisé la proc glmselect.

En partant du modèle complet :

Prixforfait = Region + Vertes + Bleues + Rouges + Noires +NbPistes + AltitudeMin + AltitudeMax + Denivele + Telesieges2places + Telesieges4places + Teleskis + Remontees

Nous avons obtenu le modèle final suivant :

Paramètres estimés				
Paramètre	**DDL**	**Estimation**	**Erreur type**	**Valeur du test t**
Intercept	1	11.737638	9.723222	1.21
NbPistes	1	0.712508	0.072548	9.82
AltitudeMin	1	0.040800	0.006843	5.96
Denivele	1	0.068941	0.005073	13.59

Table 8. Modèle complet

Plus exactement Prixforfait = 11.73 + NbPistes*0.71 + AltitudeMin*0.04 + Denivele*0.06 + Bruit.

En conclusion, notre étude nous amène à penser que les variables impactant de façon significative le prix du forfait de ski sont le nombre de pistes, l'altitude minimal et le nombre de dénivelé.

Annexe

Code SAS :

```
/* Importation des données */

proc import datafile="/home/arakchou.ayour0/sasuser.v94/forfaitski2017.csv"
   out=ski
   dbms=csv
   replace;
   delimiter=';';
   getnames=yes;


/*Premier aperçu de la base de données et sur la nature des variables*/

proc contents data=ski;
run;


/* Statistiques élémentaires */

proc means data=ski ;
var Prixforfait Vertes Bleues Rouges Noires NbPistes AltitudeMin AltitudeMax Denivele
Telesieges2places Telesieges4places Teleskis Remontees;
run;

/*Correlationx entre les var quanti*/
proc corr data=ski;
var Prixforfait Vertes Bleues Rouges Noires NbPistes AltitudeMin AltitudeMax Denivele
Telesieges2places Telesieges4places Teleskis Remontees;
run;


/*Boîte à moustaches de la variable Prixforfait*/

PROC SGPLOT DATA=ski;
  VBOX Prixforfait;
RUN ;


/* Diagramme circulaire de la variable Region */

PROC GCHART DATA =ski;
  VBAR Region;
RUN ; QUIT ;
```

```
/*Prix forfait en fonction du nbpistes*/
proc gplot data=ski;
plot Prixforfait*Nbpistes;
run;
quit;


/* On cherche à tester l'adéquation des variables quantitatives à une loi gaussienne */

proc univariate data=ski normal ;
var Prixforfait Vertes Bleues Rouges Noires NbPistes AltitudeMin AltitudeMax Denivele
Telesieges2places Telesieges4places Teleskis Remontees;
run ;


/* Boîte à moustaches de la variable Prixforfait selon la Region */

PROC SGPLOT DATA=ski;
  VBOX Prixforfait / CATEGORY=Region ;
RUN ;


/*Dichotomisation de la variable Region*/

data ski;
set ski;
if Region = 'Alpes du Nord' then D1 = 1; else D1 = 0;
if Region = 'Alpes du Sud' then D2 = 1; else D2 = 0;
if Region = 'Jura' then D3 = 1; else D3 = 0;
if Region = 'Massif central' then D4 = 1; else D4 = 0;
if Region = 'Pyrenees' then D5 = 1; else D5 = 0;
if Region = 'Vosges' then D6 = 1; else D6 = 0;
run;

/* Regression effectuee que sur les variables dichotomisées*/
proc reg data=ski;
model Prixforfait = D1 D2 D3 D4 D5 D6;
run;


/* => Modèles sur les variables caractérisant les pistes */

TITLE "Modele sur les variables de pistes";
SYMBOL1 v=diamond i=rl;
PROC REG DATA=ski;
        MODEL Prixforfait = Vertes Bleues Rouges Noires;
RUN;
QUIT;
```

```
/* sans rouge*/
PROC REG DATA=ski;
        MODEL Prixforfait = Vertes Bleues Noires;
RUN;
QUIT;


/*sans verte*/
PROC REG DATA=ski;
        MODEL Prixforfait = Bleues Noires;
RUN;
QUIT;

/* => Modèles sur les variables caractérisant l'aspect naturelle de la station*/

TITLE "Modele sur les aspects naturelles de la station";
SYMBOL1 v=diamond i=rl;
PROC REG DATA=ski;
        model Prixforfait= Region AltitudeMin AltitudeMax Denivele;
RUN;
QUIT;



/*sans AltitudeMax*/
PROC REG DATA=ski;
        model Prixforfait= AltitudeMin Denivele;
RUN;
QUIT;

/*  => Modèles sur les variables caractérisant les infrastructures de la station*/

TITLE "Modele sur les infrastructures de la station";
SYMBOL1 v=diamond i=rl;
PROC REG DATA=ski;
        model Prixforfait= Telesieges2places Telesieges4places Teleskis Remontees;
RUN;
QUIT;


/*sans teleesieges4places*/
PROC REG DATA=ski;
        model Prixforfait= Telesieges2places Teleskis Remontees;
RUN;
QUIT;
```