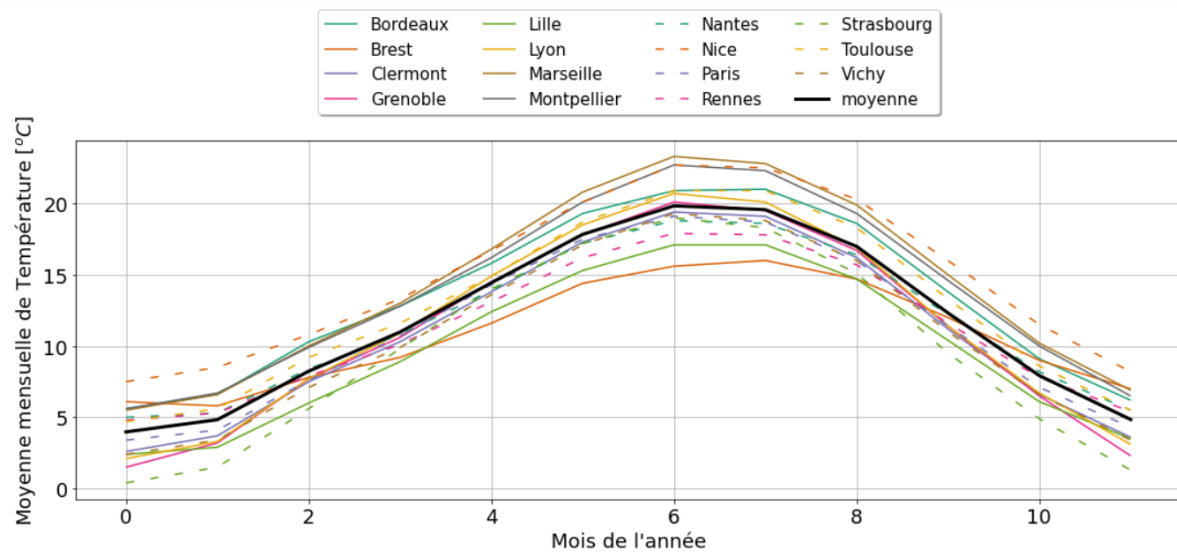


## Compte-rendu du TP n°1 – ACP

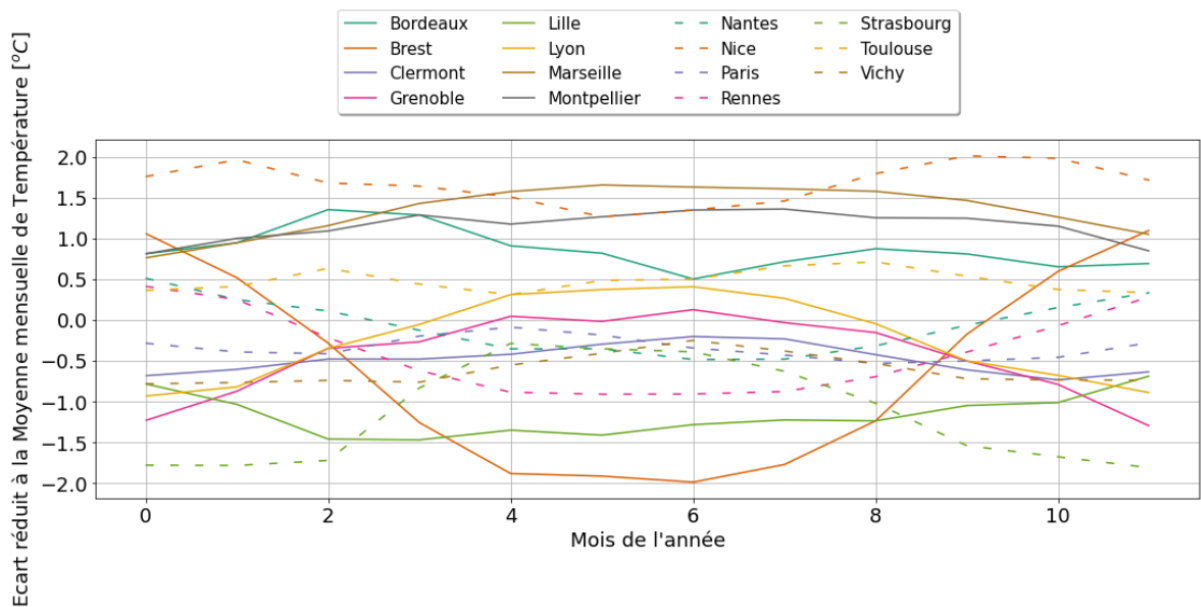
On a vu que l'ACP est l'une des techniques les plus utilisées quand on veut partir à la pêche de l'information dans de grands jeux de données : du data mining. Elle est utile lorsqu'on veut diminuer le nombre de dimensions en concentrant le maximum d'information (inertie) dans les premières dimensions (appelées axes ou facteurs).

Voici notre jeu de données, un certain nombre de moyenne de températures en fonction des mois de l'année et en fonction des différentes villes de France :



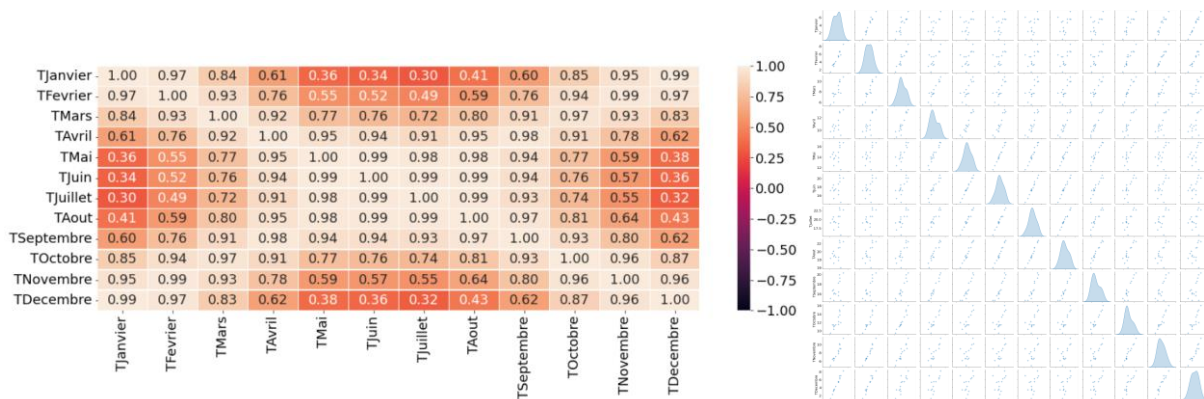
- Il n'y a rien d'anormales de particulier à constater, toutes les villes font plus ou moins la même tendance.

Une fois les données centrées et réduite :



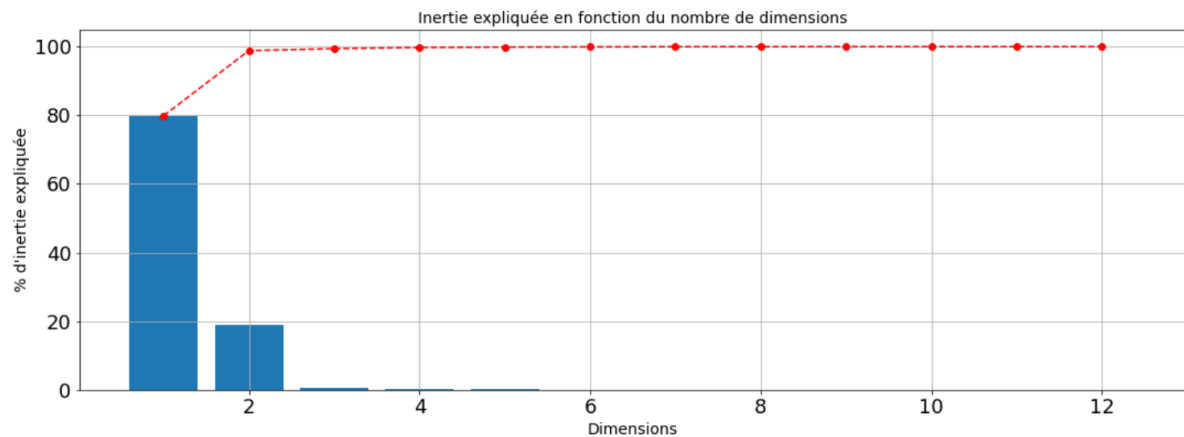
- On constate déjà des disparités avec notamment Brest qui entre Mars et Août est en dessous de la moyenne des autres villes de France (avec une forte amplitude).
- D'autres villes comme Lille où il fait globalement toute le temps la même température avec une tendance négative, la température se situe entre -5° et -1.5° tout au long de l'année.
- Et d'autres villes comme Marseille où c'est l'inverse de Lille, la température reste globalement positive entre +1° et +1.5°.

On cherche ensuite à regarder la matrice de corrélations à mettre en face avec le diagramme de dispersion :



- Dans la matrice de corrélations, on constate clairement 4 tendances qui se déportent qui représentent au final les 4 saisonnalités des températures des villes.
- Les indices de corrélations en rouges foncés se rapproche de 0.30 aux points les plus forts. Ce démontre qu'il n'y a une absence de relation linéaire mais pas qu'il y a une absence totale de corrélation. Il s'agit même d'une corrélation plutôt exponentielle/logarithmique.
- Ainsi, on voit que les mois proches sont fortement corrélés.
- **Tout cela démontre que chaque saisonnalité a sa répartition de la corrélation de température en fonction de la dispersion des mois (distance entre les mois grâce au diagramme de dispersion) et que tout cela est cyclique avec quelques oscillations de températures.**

Ensuite on a le diagramme d'inertie :



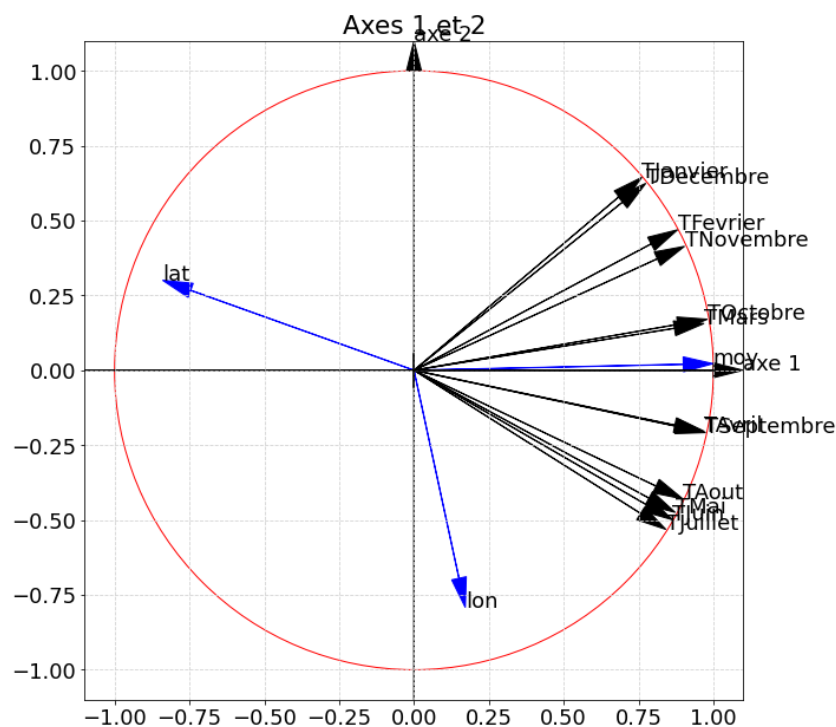
- On voit clairement 2 blocs qui représente quasiment 100% de l'information donc on va garder 2 dimensions.
- Ces 2 dimensions sont les 2 polarités des températures : chaud et froid.

On a ensuite la contribution de chaque ville par rapport à l'inertie totale :

	CP_1	CP_2	CP_3	CP_4	CP_5	CP_6	CP_7	CP_8	CP_9	CP_10	CP_11	CP_12
<b>Bordeaux</b>	94.66877	0.11612	5.05094	0.00209	0.00060	0.00021	0.15197	1.44193e-03	0.00515	2.49611e-03	5.20297e-06	1.97301e-04
<b>Brest</b>	23.43625	76.33938	0.06013	0.00751	0.04350	0.10664	0.00032	3.30715e-03	0.00002	2.59841e-03	3.39567e-04	1.66597e-06
<b>Clermont</b>	87.98844	10.37051	0.01078	0.82234	0.64317	0.02901	0.00009	7.08761e-05	0.13498	2.12152e-08	3.68729e-04	2.39268e-04
<b>Grenoble</b>	42.89404	52.25810	0.34672	2.93077	1.46948	0.03945	0.05217	2.62176e-03	0.00102	1.30178e-03	4.22474e-03	9.84728e-05
<b>Lille</b>	97.15212	1.93557	0.69198	0.00013	0.00823	0.05192	0.15049	6.89806e-06	0.00373	5.68425e-03	1.12284e-04	3.28216e-05
<b>Lyon</b>	17.81371	81.71273	0.01349	0.01310	0.07877	0.23828	0.02827	1.65328e-02	0.01926	1.54672e-04	6.54655e-02	2.43157e-04
<b>Marseille</b>	96.41953	2.83586	0.52493	0.09368	0.00785	0.08694	0.01588	1.18409e-02	0.00023	8.36727e-04	2.41117e-03	1.07076e-05
<b>Montpellier</b>	98.57584	1.08622	0.19247	0.00326	0.01029	0.00280	0.04220	7.22490e-02	0.01287	1.72315e-03	6.56042e-05	8.32836e-06
<b>Nantes</b>	5.64533	88.63242	3.71506	1.01175	0.14492	0.02361	0.23204	8.55522e-03	0.40513	1.52782e-01	2.61279e-02	2.27374e-03
<b>Nice</b>	98.00514	1.69208	0.11393	0.01216	0.09360	0.07575	0.00284	3.37285e-03	0.00050	3.99729e-06	6.03603e-04	4.07070e-06
<b>Paris</b>	88.93600	1.40945	2.74507	6.52242	0.00039	0.15267	0.00387	1.19408e-01	0.06208	4.21306e-02	3.70470e-04	6.13198e-03
<b>Rennes</b>	41.98530	56.65022	0.47064	0.01330	0.21569	0.33190	0.28807	2.52355e-02	0.00831	1.09338e-02	3.39596e-04	6.46092e-05
<b>Strasbourg</b>	77.56541	21.71378	0.00550	0.63765	0.03710	0.01444	0.00068	1.39610e-02	0.00508	6.13770e-03	2.65086e-04	6.30970e-08
<b>Toulouse</b>	95.25552	0.58559	0.05945	2.82812	0.75989	0.07641	0.25992	1.50109e-02	0.15591	1.48005e-03	1.11722e-05	2.66705e-03
<b>Vichy</b>	92.15064	6.29104	0.80485	0.40569	0.22151	0.00001	0.10243	1.66453e-02	0.00189	2.68161e-03	2.32154e-03	2.93217e-04

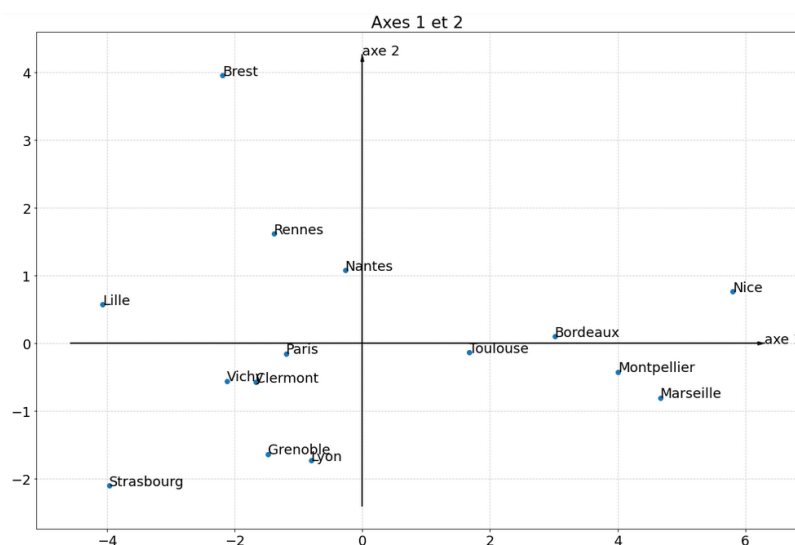
- Le constat précédent est vérifié, on voit bien 2 blocs/composants qui contiennent toute l'information et cela n'est pas répartie toujours équitablement.
- On voit d'ailleurs que certaines villes ont besoin de ces 2 polarités pour en tirer l'ensemble des informations, exemple de Grenoble où on a une répartition 42/52 (qui représente le chaud/froid).
- D'autres villes où on a plus de poids le 2<sup>ème</sup> bloc comme Brest avec 23/76 (on peut donc imaginer que le 2<sup>nd</sup> bloc c'est le bloc du froid).
- Et la plupart des villes ont une répartition forte sur le 1<sup>er</sup> bloc.

On a ensuite le cercle de corrélation (avec teste des hypothèses avec longitude et latitude) :



- Il y a un lien entre la dispersion des saisons (axe 2) et les températures (axe 1).
- On constate ainsi que les températures d'une manière générale pour Décembre et Janvier sont très similaires, et que Mai, Juin, Juillet et Août sont aussi très similaires et qu'il y a une corrélation avec la dispersion des saisons puisque ces différents mois représentent bien les saisons hiver et été. D'ailleurs, on constate presque un angle droit entre l'axe Décembre/Janvier et Mai/Juin/Jullet/Août qui sont contraires (presque présence d'un angle droit).

Enfin, si on met en perspective le cercle de corrélation et le graphique des nuages des individus :



- On constate que ce graphique des nuages d'individus représente la carte de la France de manière retournée horizontalement et verticalement.
- On constate alors que les 2 hypothèses de latitude (dans le cercle de corrélation) et la moyenne des températures (axe 1) sont liées. Par exemple Lille est une ville très élevée en latitude (donc au Nord), on voit qu'elle se situe très à gauche du graphique donc des températures très basses. Contrairement à Nice ou c'est une ville plus basse (Sud) en latitude, on voit que c'est une ville où il faut beaucoup plus chaud.
- D'ailleurs, on voit que sur l'axe 2, Brest est très élevée et cela montre qu'il y a une grosse dispersion des saisons par rapport à la moyenne et on avait déjà fait ce constat sur le premier graphique des températures une fois centrées et réduites (cela ne fait que confirmer ce constat).
- Aussi, l'autre hypothèse concernant la longitude (toujours dans le cercle de corrélation) et la moyenne des températures est toujours vérifiées (axe 1) car les villes du Sud de la France sont très concentrées autour de l'axe des longitudes. Ainsi, plus on va vers le Sud, plus la longitude n'a pas d'impact sur la moyenne des températures contrairement au Nord.
- **Il faut tout de même noter quelques cas particuliers comme Grenoble et Lyon qui sont plutôt dans le Sud mais où ici les températures sont plutôt basses (ce qui n'est pas normal par rapport à notre constat).**