BAZOGE Adrien
BRASSIER Maëlle
CATELLA Solène
LOIZON Maxime

M1 - ATAL

BASE DE DONNÉES ÉVOLUÉES

Rapport de projet

Emploi du temps journalier d'un individu

SOMMAIRE

1. Présentation du sujet

a. Choix du sujet

Dans le cadre de ce projet, nous nous sommes intéressés à **l'emploi du temps journalier d'un individu, à partir du temps moyen passé sur un ensemble d'activités**. L'objectif ici visé était de pouvoir identifier, classer voire quantifier les principaux types d'activités auxquelles un citoyen se livre au cours d'une période donnée (ici année), pour un type de profil donné (sexe, situation maritale, niveau d'études, situation professionnelle, etc.).

b. Datasets utilisés

Les datasets sur lesquels nous avons travaillé ont été récupérés et construits à partir du site Web *mtusdata.org* (Multinational Time Use Study). Cette plateforme met à disposition des données de recensement et d'enquête du monde entier, intégrées dans le temps (sur cinq décennies) et dans l'espace (pour 25 pays différents) :

SAMPLE DESCRIPTION						
Sample	Sample size	Age Range	Multiple respondents	Multiple days	With whom	Secondary activities
Austria, 1992	4,777	14+	Yes	No	Yes	Yes
Bulgaria, 2001	14,714	10+	Yes	Yes	Yes	Limited
Canada, 2010	15,390	15+	No	No	Yes	Yes
Finland, 1979	12,057	10+	No	Yes	Yes	Limited
Finland, 2009	7,455	10+	Yes	Yes	Yes	Detailed
France, 1985	16,047	15+	Yes	No	Yes	Yes
France, 1998	15,441	15+	Yes	No	Yes	Yes
Hungary, 1999	10,983	15+	No	Yes	Yes	Yes
Hungary, 2009	8,390	10+	No	No	Yes	Yes
Israel, 1991	4,843	14+	Yes	Yes	Yes	No
Italy, 2002	51,206	3+	Yes	No	Yes	Detailed
Netherlands, 1975	9,163	12-98	No	Yes	No	Limited
Netherlands, 1985	22,841	12-91	No	Yes	No	Limited
Netherlands, 1990	23,905	12-90	No	Yes	No	Limited
Netherlands, 2000	12,691	11-99	No	Yes	No	Yes
Netherlands, 2005	15,428	12+	No	Yes	No	Yes
Spain, 2009	19,295	10-80	Yes	No	Yes	Yes
United Kingdom, 1974-75	20,252	5-80	Yes	Yes	No	Yes

Figure 1. *Caractéristiques des échantillons MTUS (extrait)*

Échantillons

Par souci d'homogénéisation et de volumétrie, tous les échantillons n'ont pas été retenus.

Pour faciliter la recherche comparative et l'analyse, **les échantillons disponibles sur les mêmes années ont dans un premier temps été récupérés. Les échantillons en exemplaire unique** (cf. Autriche, Canada, Italie, etc.), bien que souvent disponibles sur des années non communes, **ont toutefois été conservés** : nous avons en effet jugé préférable de ne pas les écarter, de façon à ne pas trop restreindre nos analyses. Aussi, pour pouvoir mesurer l'évolution du temps attribué à une activité donnée sur plusieurs années, **les échantillons échelonnés sur des périodes de temps régulières ont été privilégiés.** C'est le cas des Pays-Bas ou des États-Unis, où il est possible de mesurer une évolution temporelle par tranche de 10 années, sur une période de 30 ans.

Un générateur d'extraits de données, mis à disposition sur le site (MTUS-X), nous a permis de mener à bien cette présélection, sous format CSV. Les échantillons retenus sont listés dans le tableau récapitulatif, joint ci-dessous :

Échantillon	Taille de l'échantillon (nombre de lignes)	Années	Part de représentativité (%)
Autriche	25 233	1992	6.2
Canada	15 390	2000	3.7
Finlande	19 493	1979, 2009	4.8
France	31 489	1985, 1998	7.7
Hongrie	8 391	2009	2.1
Israël	4 843	1991	1.2
Italie	51 206	2002	12.6
Pays-Bas	70 021	1975, 1985, 1995, 2005	17.2
Espagne	66 069	2002, 2009	16.2
Royaume-Uni	37 558	1974, 1983, 1995, 2005	9.2
États-Unis	78 000	1965, 1975, 1985, 1995, 2005	19.1
TOTAL	**407 693**	**/**	100

NB : Un tuple correspond à **l'agrégation de la journée type d'une personne, pour un ensemble d'activités données**.

Attributs

Au niveau des attributs, on distingue d'une part (1) les **attributs temporels**, liés aux mesures de temps sur un ensemble d'activités, et d'autre part (2) les **attributs relatifs aux caractéristiques personnelles des individus sondés.**

(1) Les mesures de temps ont été relevées sur les activités suivantes :

- Soins apportés aux nourrissons et enfants (*Childcare*)
- Cuisine (*Cooking*)
- Repas (*Meal*)
- Tâches ménagères (*Housework*)
- Courses alimentaires (*Grocery*)
- Soins personnels (*Washing & dressing*)
- Repos (*Sleep*)
- Études (*School*)
- Travail (*Work*)
- Pause travail (*Work break*)
- Loisirs (*Home Leisure*)
- Médias (*Media*)
- Pratique(s) religieuse(s) (*Religion*)
- Sport (*Sport*)
- Voyage (*Travel*)

(2) Les attributs intrinsèques aux individus sondés se déclinent en 4 grands ensembles :

- **Renseignements techniques** (identifiant du sujet interrogé, pays ou région de l'enquête)
- **Caractéristiques personnelles et démographiques** (âge, sexe, citoyenneté, situation maritale, état de santé, niveau d'études)
- **Statut professionnel** (poste occupé, secteur d'emploi, revenu, étudiant ou non)
- **Caractéristiques du ménage** (taille du foyer, âge du plus jeune enfant, zone habitée, possession d'un véhicule ou d'un ordinateur)

Ces attributs seront plus amplement détaillés dans la suite de ce rapport (cf. Annexe).

2. Choix de modélisation et de conception

a. Architecture de l'entrepôt de données

Concernant la modélisation de notre entrepôt de données, notre choix s'est porté sur le **modèle relationnel**. Outre le fait que les SGBDR facilitent la gestion et l'entretien du stockage des données (propriétés ACID), ils ont également l'avantage de bien se prêter à la **modélisation de données structurées**. Or, dans la mesure où les données d'enquête ont préalablement fait l'objet d'une harmonisation[1], la définition d'un **schéma fixe** s'avérait concordant. Aussi, ce type de modèle fournit une **bonne représentation de la réalité**, et permet donc de mieux apprécier les analogies avec le monde réel.

Ces différents critères nous ont donc amenés à opter pour l'architecture ROLAP. La modélisation sous forme de **schéma en étoile** est visualisable ci-contre :

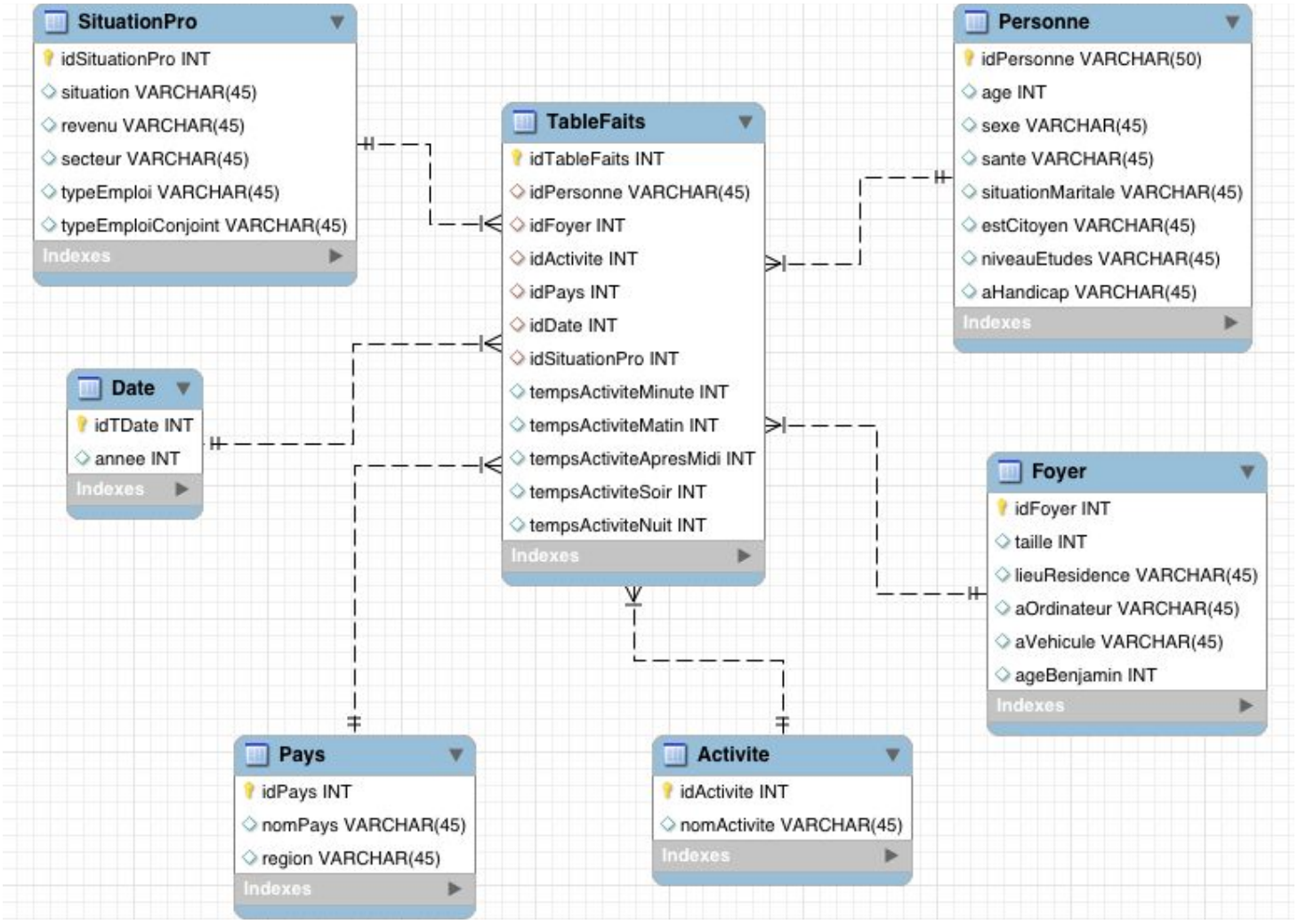


Figure 2. *Schéma en étoile de notre entrepôt de données*

[1] Le projet IPUMS est à l'origine de cette initiative, l'idée étant d'assurer la compatibilité des données dans le temps et l'espace.

b. Sélection du processus d'entreprise modélisé et déclaration du grain

La conception d'un entrepôt de données débute nécessairement par une phase de sélection du processus d'entreprise à modéliser. De par la nature même des données et les raisons qui ont motivé leur collecte, nous nous sommes ici intéressés au habitudes de vie d'un individu, et ce au travers du **temps passé sur un ensemble d'activités et de tâches données, dans un contexte situationnel, temporel et spatial déterminé.**

De façon à garantir un niveau de détail aussi fin que possible et ainsi rendre les analyses plus souples, le temps alloué à une activité/tâche donnée a été subdivisé en 4 plages horaires de 6 heures :

- **le matin (de 06h à 12h) ;**
- **l'après-midi (de 12h à 18h)**
- **le soir (de 18h à minuit)**
- **la nuit (de minuit à 6h)**

L'extracteur MTUS-X offre en effet la possibilité de créer des variables temporelles pour une activité donnée et selon une plage horaire spécifiée.

c. Table des faits

Dans ce modèle, **un fait correspond à la durée d'une activité**. La table de faits comporte tous les faits numériques nécessaires à l'analyse, ici exprimés en minutes.

Le détail des attributs de cette table est disponible en Annexe 1 de ce rapport.

d. Tables de dimensions

Les tables de dimensions, ici au nombre de six, permettent de donner un sens aux données stockées dans la table des faits, selon différents axes d'analyse.

Les propriétés des tables de dimensions ***Personne*, *Activité*, *Foyer*, *Pays*, *Date* et *Situation professionnelle*** sont détaillées en Annexe 2 de ce rapport.

3. Construction de l'entrepôt de données et difficultés rencontrées

a. Construction de l'entrepôt

La construction de l'entrepôt de données a nécessité un long travail de préparation des données.

i. Sélection et extraction des données

Une première étape a consisté à **sélectionner et extraire les données** via le générateur MTUS-X. Cet outil offre en effet la possibilité de construire manuellement son propre jeu de données, en choisissant un à un les différents attributs (indicateurs d'analyse) et enregistrements (échantillons) désirés. La sélection s'est donc faite à deux niveaux :

1. choix des échantillons ;
2. choix des indicateurs d'analyse/attributs.

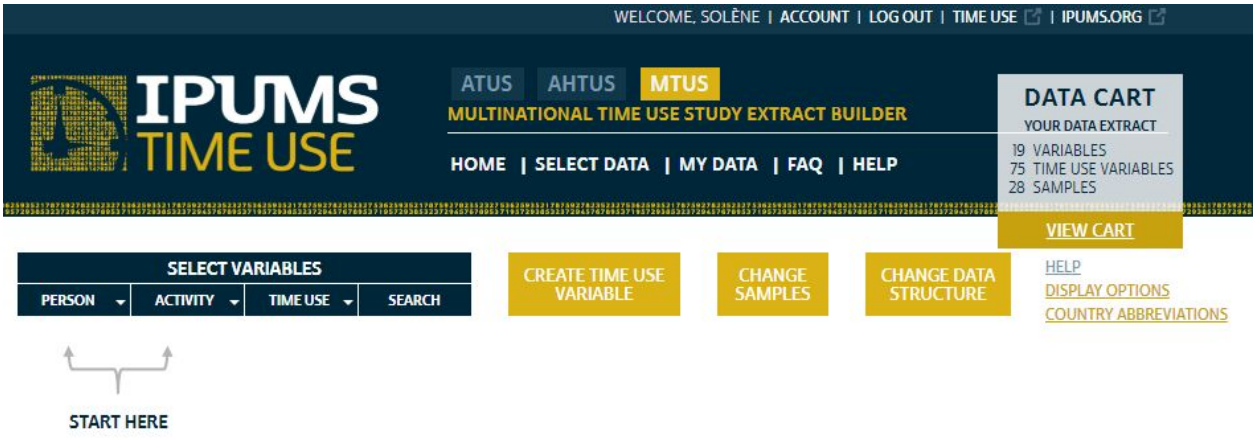


Figure 3. *Interface principale de MTUS-X*

La sélection des échantillons a déjà été discutée plus haut (cf. page 4). Ainsi, sur 40 échantillons, 25 ont été conservés, de façon à concevoir un jeu de données aussi homogène que possible.

Le processus de sélection relatif aux indicateurs d'analyse a quant à lui demandé plus de temps. Le niveau de granularité associé aux données renseignées étant très fin, il a fallu trier et distinguer, parmi tous les critères de mesure possibles, ceux présentant le plus d'intérêt. Cette sélection s'est faite sur la base du contenu descriptif de chaque attribut, renseigné par l'extracteur (voir ci-contre).

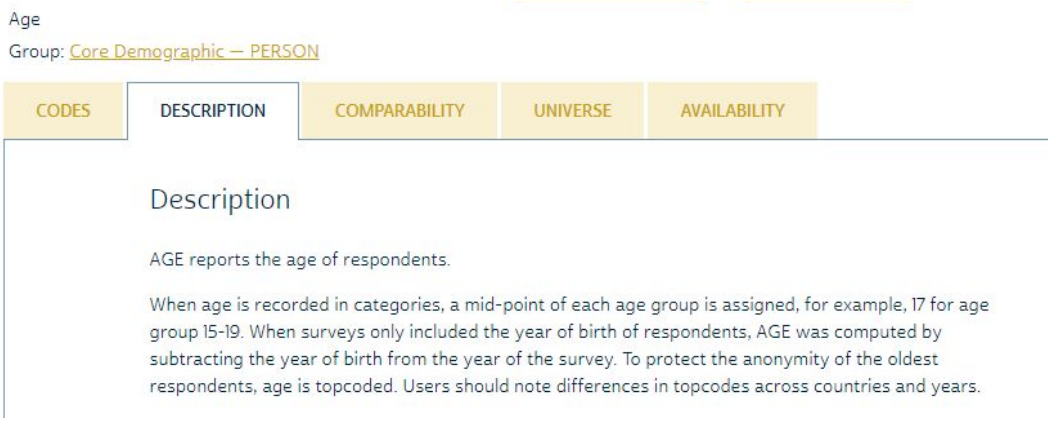


Figure 4. *Description de l'attribut "Age"*

Sur un total de 56 attributs, 22 ont été retenus (hors attributs temporels). Il s'agit des attributs précédemment cités, contenus dans les tables des faits et dimensions.

Les attributs temporels ont quant à eux été créés un à un, pour chacune des différentes activitées évaluées. Ainsi, comme mentionné précédemment, une activité à été mesurée sur 4 plages horaires distinctes (matin, après-midi, soir, nuit), auquel il faut ajouter un indicateur de mesure temporel global (temps consacré sur la journée entière)

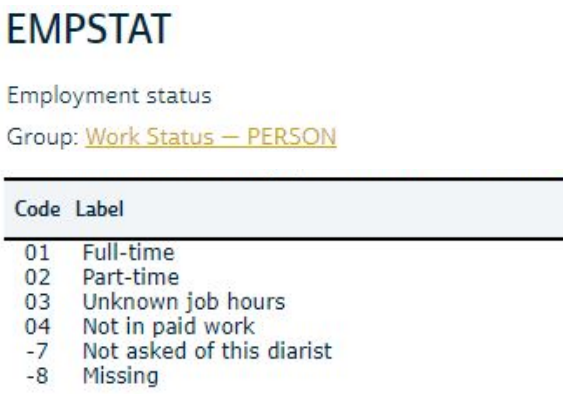
Time Use Variable	Label
ChildCare	ACT: Child care
ChildCareAfternoon	ChildCareAfternoon
ChildCareEvening	ChildCareEvening
ChildCareMorning	ChildCareMorning
ChildCareNight	ChildCareNight
Cooking	ACT: Unpaid domestic work
CookingAfternoon	CookingAfternoon
CookingEvening	CookingEvening
CookingMorning	CookingMorning
CookingNight	CookingNight
Grocery	ACT: Unpaid domestic work
GroceryAfternoon	GroceryAfternoon
GroceryEvening	GroceryEvening
GroceryMorning	GroceryMorning
GroceryNight	GroceryNight
HomeLeisure	ACT: In home free time leisure
HomeLeisureAfternoon	HomeLeisureAfternoon
HomeLeisureAfternoonEvening	HomeLeisureAfternoonEvening
HomeLeisureAfternoonNight	HomeLeisureAfternoonNight
HomeLeisureMorning	HomeLeisureMorning

Figure 5. *Extrait des attributs temporels*

ii. Nettoyage et intégration des données

La phase de nettoyage et d'intégration des données s'est avérée particulièrement chronophage.

Une première étape a consisté à convertir sous forme textuelle et descriptive l'ensemble des valeurs prises par chaque attribut catégorique. Dans le jeu de données extrait, ces différentes valeurs étaient en effet associées à une valeur d'index arbitraire, sous forme de "code" (cf. figure 6). Afin d'accroître la sémantique des données et en faciliter l'interprétation, une mise en correspondance a été effectuée entre le code d'un attribut donné et sa valeur descriptive, tel que présenté ci-contre :



EMPSTAT

Employment status

Group: Work Status — PERSON

Code	Label
01	Full-time
02	Part-time
03	Unknown job hours
04	Not in paid work
-7	Not asked of this diarist
-8	Missing

Figure 6. *Table de correspondance de l'attribut EMPSTAT*

L'étiquetage a été effectué à partir d'un fichier JSON, via un script Python. Ce traitement a été appliqué sur l'ensemble des attributs catégoriques du dataset, à savoir nomPays (COUNTRY), lieuResidence (URBAN), sexe (SEX), estCitoyen (CITIZEN), aVehicule (VEHICLE), aOrdinateur (COMPUTER), secteur (SECTOR), sante (HEALTH), aHandicap (DISAB), revenu (INCOME) et niveauEtudes (EDUCA).

```
"URBAN":
{
    "01" : "Citadin",
    "02" : "Rural",
    "-7" : "Pas applicable",
    "-8" : "N/A"
},
"EMPSTAT":
{
    "01" : "Temps plein",
    "02" : "Temps partiel",
    "03" : "Heures d'emploi inconnues",
    "04" : "Sans emploi rémunéré",
    "-7" : "Pas demandé",
    "-8" : "N/A"
},
```

Figure 7. *Extrait du fichier json de mapping général*

Des difficultés ont été rencontrées au niveau du décodage de l'attribut EDUCA (niveau d'études). Chaque pays ayant son propre système éducatif, les tables d'association index/valeur ne concordaient pas toujours. Un mapping a donc été nécessaire pour chacun des différents pays présents dans la base, tel que présenté ci-contre :


```
{
    "AT":
    {
        "1" : "Elémentaire",
        "3" : "Cycle professionnel",
        "4" : "Supérieur général",
        "5" : "Supérieur professionnel",
        "6" : "Université",
        "7" : "Enfants âgés entre 10 et 14 ans"
    },
    "CA" :
    {
        "1" : "Doctorat/Master",
        "3" : "Baccalauréat",
        "4" : "Diplôme/certificat d'un collège communautaire",
        "5" : "Diplôme/certificat d'une école de métiers ou d'une école technique",
        "6" : "Université",
        "7" : "Diplôme/certificat professionnel autre",
        "8" : "Diplôme d'études secondaires",
        "9" : "Diplôme d'études secondaires autre",
        "10" : "Elémentaire/Non scolarisé"
    },
```


Figure 8. *Extrait du fichier json de mapping de l'attribut EDUCA*

Après réflexion cependant, nous avons préféré ne pas harmoniser cet attribut, et ce pour deux raisons principales :

- d'abord parce qu'il aurait été trop compliqué d'unifier à la main et sous un même standard ces différents systèmes éducatifs ;
- ensuite parce que l'attribut EDTRY, indiquant le niveau d'éducation le plus élevé du répondant d'après la Classification internationale type de l'éducation (ISCED), répondait déjà à cet objectif.

L'attribut EDUCA a donc été substitué par l'attribut EDTRY.

Les prétraitements se sont poursuivis par **la gestion des valeurs manquantes, déclinées sous trois variantes possibles** :

- donnée non renseignée (*not asked*) ;
- donnée indiquée comme "manquante" (*missing*) ;
- donnée n'ayant pas pu être enregistrée par absence de champ (*could not be created*).

En effet, bien que ce type d'enquête soit standardisé, il ne faut pas oublier que les données ici récoltées l'ont été sur plusieurs décennies et au travers de multiples pays. Certains critères de mesure ont donc pu disparaître, voire évoluer. Ces trois variantes possibles ont été traduites dans la base de données comme étant respectivement "N/A", "Pas demandé", ou "Pas applicable".

Une troisième étape a consisté à créer un attribut agrégat par personne, par activité. En effet, certains échantillons de données comportaient plusieurs entrées pour un même individu (étude du temps passé pour une activité sur plusieurs jours). Une mesure moyenne du temps passé par activité par personne a donc été insérée dans la base de données, en tant qu'attribut agrégat. Pour ce faire, il a été nécessaire de créer un nouvel identifiant par individu sondé, car l'attribut d'origine IDENT référençait parfois deux individus sur une échelle spatiale et temporelle différente comme une seule et même personne (confrontation de données syntaxiquement semblables mais sémantiquement différentes). Pour ce faire, nous avons utilisé Talend Data Preparation, un outil facilitant la gestion et la préparation de données.

Enfin, un attribut *region* a été créé, et ce afin de regrouper sous un même nom les pays faisant partie d'un même continent (Europe, Amérique du Nord). Cet indicateur global géographique nous permettra de faire des analyses à un niveau de granularité plus élevé.

b. Chargement des données dans l'entrepôt

Afin de charger les données dans l'entrepôt, **une première étape a consisté à regrouper l'ensemble des données issues des différents CSV en 7 tableaux distincts à l'aide d'un script Python**, ces tableaux représentant les tables des faits et dimensions précédemment mentionnées. Les fichiers les plus volumineux ont quant à eux été séparés en plusieurs CSV de 8000 lignes. Cette gestion séparée des fichiers et des tables nous a permis d'insérer nos données plus facilement dans la base, à partir de script SQL. Pour les États-Unis par exemple, 78 fichiers d'insertion SQL au total ont été générés.

Il avait été décidé en amont que le déploiement de la base de données se ferait sous *MySQL*, en utilisant *Wamp* et *phpMyAdmin*. Cependant, dans la mesure où ce système de gestion de base de données ne supporte pas les requêtes OLAP, **nous avons finalement opté pour *Oracle Express (XE)***. Le passage sous *Oracle* a nécessité quelques modifications sur les fichiers SQL d'origine :

- les doubles guillemets ont dû être substitués par des guillemets simples ;
- la version d'Oracle utilisée ne prenant pas en charge l'auto incrément, un trigger a dû être créé, de façon à incrémenter l'attribut idTableFaits via une séquence.

4. Requêtage de l'entrepôt de données

Les requêtes effectuées nous ont tout aussi bien amenés à repérer des habitudes de vie par pays, mais aussi à mesurer des évolutions temporelles au sein sein d'un même pays, sur plusieurs années.

Requête 1 (opérateurs GROUP BY, RANK)

Palmarès des 5 premiers pays qui consacrent le plus de temps au travail

```
Select *  FROM (SELECT pa.nomPays AS Pays, ac.nomActivite,
AVG(t.tempsActiviteMinute) AS Duree
FROM tableFaits t,  Activite ac, Pays pa
WHERE t.idPays = pa.idPays
AND t.idActivite = ac.idActivite
AND (ac.nomActivite = 'Work')
GROUP BY pa.nomPays, ac.nomActivite
ORDER BY Duree DESC)
WHERE ROWNUM <= 5;
```

	Pays	Duree
1	Espagne	7h 30 min
2	Canada	7h 28 min
3	Israël	7h 14 min
4	Italie	7h 13 min
5	États-Unis	7h 02 min

Les Espagnols sont donc les plus travailleurs, avec une moyenne de 7h30 minutes par jour.

Requête 2 (opérateurs GROUP BY, RANK)

Top 5 des activités sur lesquelles un individu passe le plus de temps sur une journée

```
SELECT * FROM (SELECT a.nomActivite
AS Activite, ROUND(AVG(t.tempsActiviteMinute)/60,0) AS Duree_heure
FROM TableFaits t, Pays pa, tDate d, Activite a
WHERE t.idPays = pa.idPays
AND t.idDate = d.idDate
AND t.idActivite = a.idActivite
GROUP BY a.NomActivite
ORDER BY Duree_heure DESC)
WHERE ROWNUM <=5;
```

	Activite	Duree_heure
1	Sleep	8
2	Work	7
3	School	5
4	Media	3
5	Housework	3

L'activité la plus chronophage est le repos avec une moyenne de 8 heures par jour, talonné de près par le travail avec 7 heures. S'en suit le temps consacré aux études de 5 heures et enfin ceux aux médias et tâches ménagères, tous les deux égaux de 3 heures.

Requête 3 (opérateur GROUP BY CUBE)

Temps moyen consacré aux tâches ménagères par pays et par sexe

```
SELECT pa.nomPays AS Pays, pe.Sexe AS Sexe,
ROUND(AVG(t.tempsActiviteMinute),0) AS Temps
FROM TableFaits t, Pays pa, tDate d,
Activite a, Personne pe
WHERE t.idPays = pa.idPays
AND t.idDate = d.idDate
AND t.idActivite = a.idActivite
AND t.idPersonne = pe.idPersonne
AND a.nomActivite = 'Housework'
GROUP BY CUBE (pa.nomPays, pe.Sexe)
```

Pays	Sexe	Temps
		189
	Femme	233
	Homme	118
Canada		175
Canada	Femme	190
Canada	Homme	153
France		163
France	Femme	204

Pays	Sexe	Temps
Italie		202
Italie	Femme	263
Italie	Homme	102
Espagne	Femme	247
Espagne	Homme	107
Hongrie		184
Hongrie	Femme	225
Hongrie	Homme	122

Pays	Sexe	Temps
Finlande		150
Finlande	Femme	178
Finlande	Homme	115
Pays-Bas		178
Pays-Bas	Femme	212
Pays-Bas	Homme	124
RU		182
RU	Femme	224

France	Homme	105
Israël		172
Israël	Femme	206
Israël	Homme	99

Autriche		232
Autriche	Femme	277
Autriche	Homme	146

RU	Homme	111
Etats-Unis		151
Etats-Unis	Femme	171
Etats-Unis	Homme	120

De façon générale, les femmes passent plus de temps sur les tâches ménagères que les hommes. Israël apparaît comme étant le pays le plus inégalitaire dans l'attribution des tâches ménagères entre hommes et femmes. Avec une répartition quasiment égale du temps alloué, la Finlande se positionne en tête de classement.

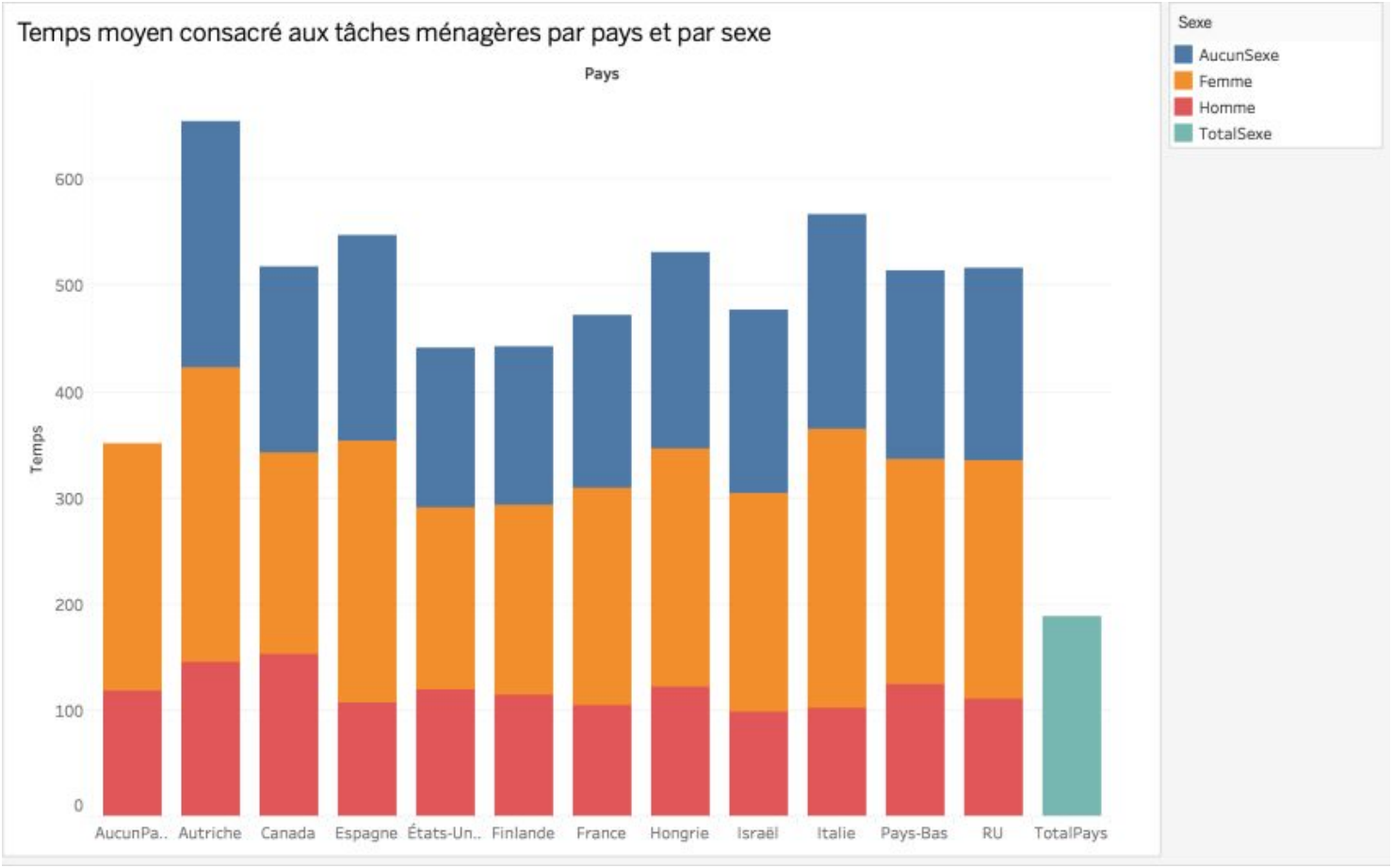


Figure 9. *Graphique représentant les résultats de la requête 3*

Requête 4 (opérateur GROUP BY)

Palmarès des pays consacrant le plus de temps aux médias (ordinateur, télévision) le soir, à compter des années 2000

```
SELECT p.nomPays as Pays, ROUND(AVG(t.tempsActiviteSoir),0) AS Temps_min
FROM Pays p, tableFaits t, Activite a, tDate d
WHERE t.idDate = d.idDate
AND t.idActivite = a.idActivite
AND t.idActivite = a.idActivite
AND t.idPays = p.idPays
AND a.nomActivite = 'Media'
AND d.annee >= 2000
GROUP BY p.nomPays ORDER BY Temps_min DESC;
```

Pays	Temps_min
Finlande	131
Hongrie	131
Canada	131
États-Unis	128
Pays-Bas	117
Espagne	101

La Finlande, la Hongrie et le Canada, suivis de près par les États-Unis, sont les pays où les individus consacrent le plus de temps aux médias le soir.

Requête 5 (opérateur GROUP BY GROUPING SETS)

Évolution du temps consacré aux activités *Travail* et *Repos* aux Pays-Bas entre 1975 et 2005

```
SELECT a.nomActivite AS Activite, d.annee AS Annee,
ROUND(AVG(t.tempsActiviteMinute),0) AS Temps
FROM TableFaits t, Pays pa, tDate d, Activite a
WHERE t.idPays = pa.idPays AND t.idDate = d.idDate
AND t.idActivite = a.idActivite
AND (a.nomActivite = 'Work' OR a.nomActivite = 'Sleep')
AND pa.nomPays = 'Pays-Bas' AND (d.annee = 1975 OR d.annee = 2005)
GROUP BY GROUPING SETS ((d.annee,a.nomActivite), (a.nomActivite));
```

Activite	Annee	Temps
Work	1975	382
Work	2005	403
Work		397
Sleep	1975	516
Sleep	2005	518
Sleep		517

Entre 1975 et 2005, c'est-à-dire sur une période de 30 ans, le temps alloué par les Hollandais au travail a grimpé de 15 minutes, pour une durée de sommeil équivalente.

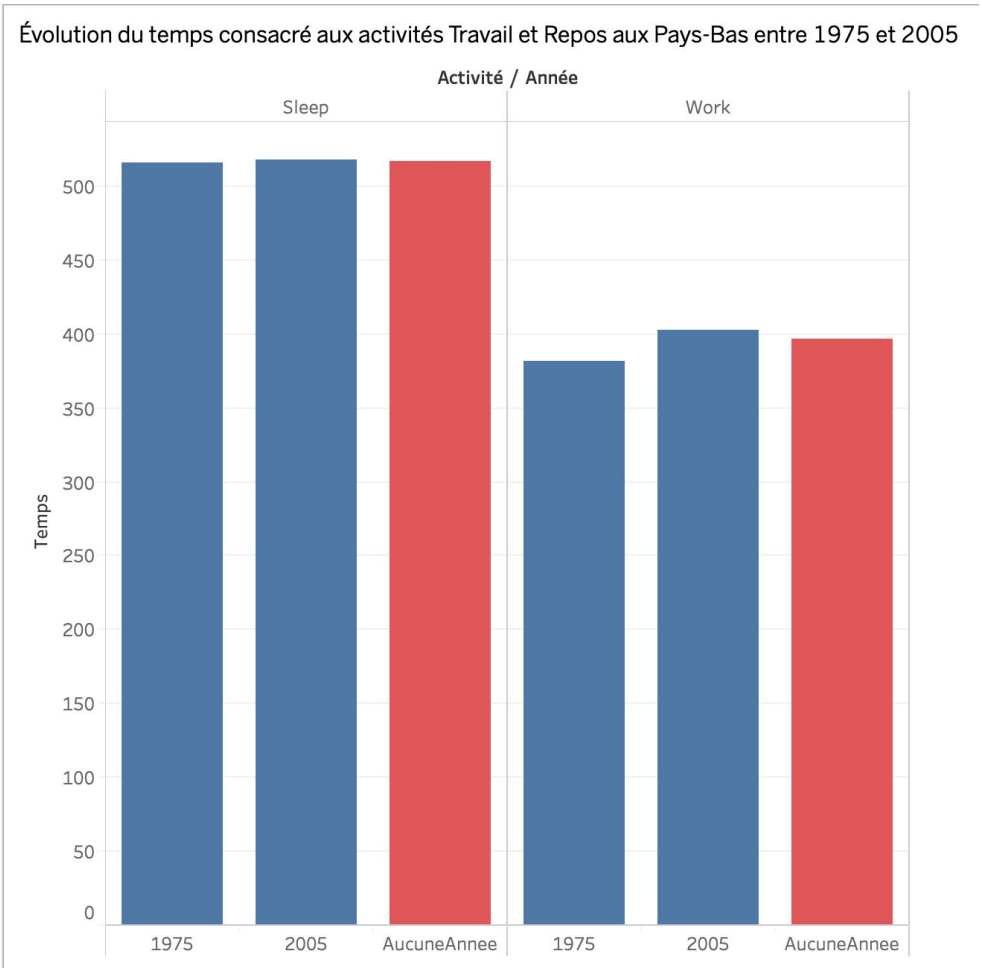


Figure 10. *Graphique représentant les résultats de la requête 5*

Requête 6 (opérateur GROUP BY ROLLUP)

Temps moyen consacré aux activités de type *Loisirs* pour une situation maritale et une zone d'habitation données

```
SELECT pe.situationMaritale AS Situation_Maritale, f.lieuResidence AS
Residence, AVG(t.tempsActiviteMatin) AS Matin,
AVG(t.tempsActiviteApresMidi) AS Après-Midi,
AVG(t.tempsActiviteSoir) AS Soir,
AVG(t.tempsActiviteNuit) AS Nuit,
FROM TableFaits t, Activite a, Personne pe, Foyer f
WHERE t.idActivite = a.idActivite
AND t.idPersonne = pe.idPersonne
AND  t.idFoyer = f.idFoyer
AND a.nomActivite = 'homeLeisure'
GROUP BY ROLLUP (pe.situationMaritale, f.lieuResidence);
```

Situation_Maritale	Residence	Matin	Après-Midi	Soir	Nuit
	Rural	21	67	31	2
	Citadin	18	67	31	2
		19	67	31	2
En couple	Rural	15	52	58	6
En couple	Citadin	14	49	57	7
En couple		15	50	57	7
Célibataire	Rural	17	60	60	4
Célibataire	Citadin	16	57	62	7
Célibataire		16	58	61	6

D'après les résultats obtenus, il semblerait que les célibataires ruraux sortent davantage que les individus en couple, quel que soit le statut marital de ces derniers.

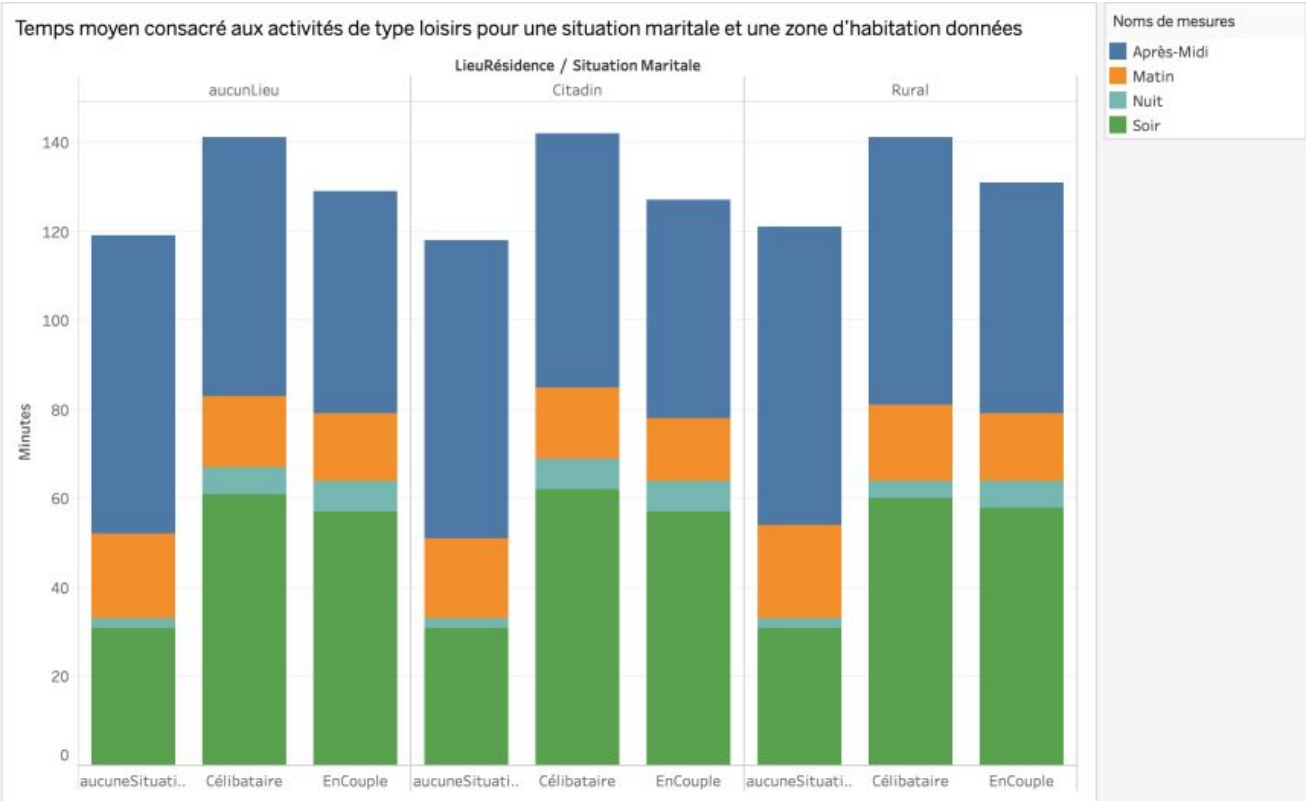


Figure 11. *Graphique représentant les résultats de la requête 6*

Requête 7 (opérateurs PARTITION BY, DENSE_RANK, WHEN/BETWEEN/THEN)

Temps moyen consacré aux activités *Média*, *Sport* et *Voyage* par tranche d'âge

```
SELECT a.nomActivite AS Activite, CASE
WHEN pe.age <= 10 THEN '1-10'
WHEN pe.age BETWEEN 11 and 20 THEN '11-20'
WHEN pe.age BETWEEN 21 and 30 THEN '21-30'
WHEN pe.age BETWEEN 31 and 40 THEN '31-40'
WHEN pe.age BETWEEN 41 and 50 THEN '41-50'
WHEN pe.age BETWEEN 51 and 60 THEN '51-60'
WHEN pe.age BETWEEN 61 and 70 THEN '61-70'
WHEN pe.age BETWEEN 71 and 80 THEN '71-80'
WHEN pe.age BETWEEN 81 and 90 THEN '81-90' ELSE '90+' END AS Age,
ROUND(AVG(t.tempsActiviteMinute)), dense_rank() OVER (PARTITION BY
a.nomActivite ORDER BY AVG(t.tempsActiviteMinute) DESC) rank AS Rang
FROM tableFaits t, Personne pe, Activite a
WHERE a.idActivite = t.idActivite AND pe.idPersonne = t.idPersonne
AND pe.age != '-8' AND (a.nomActivite = 'Sport' OR a.nomActivite =
'Travel' OR a.nomActivite = 'Media')
```

```
GROUP BY a.nomActivite, CASE WHEN pe.age <= 10 THEN '1-10'
WHEN pe.age BETWEEN 11 and 20 THEN '11-20'
WHEN pe.age BETWEEN 21 and 30 THEN '21-30'
WHEN pe.age BETWEEN 31 and 40 THEN '31-40'
WHEN pe.age BETWEEN 41 and 50 THEN '41-50'
WHEN pe.age BETWEEN 51 and 60 THEN '51-60'
WHEN pe.age BETWEEN 61 and 70 THEN '61-70'
WHEN pe.age BETWEEN 71 and 80 THEN '71-80'
WHEN pe.age BETWEEN 81 and 90 THEN '81-90' ELSE '90+' END;
```

Activite	Age	Temps	Rang
Media	71-80	278	1
Media	81-90	278	2
Media	90+	267	3
Media	61-70	244	4
Media	51-60	203	5
Media	11-20	190	6
Media	41-50	176	7
Media	21-30	165	8
Media	31-40	159	9
Media	1-10	154	10
Sport	61-70	125	1
Sport	11-20	124	2
Sport	81-90	119	3
Sport	71-80	118	4
Sport	1-10	117	5
Sport	51-60	103	6
Sport	51-60	111	6
Sport	41-50	103	2

Activite	Age	Temps	Rang
Sport	41-50	103	7
Sport	21-30	100	8
Sport	31-40	97	9
Sport	90+	27	10
Travel	21-30	92	1
Travel	41-50	89	2
Travel	31-40	89	3
Travel	11-20	87	4
Travel	51-60	83	5
Travel	61-70	76	6
Travel	71-80	68	7
Travel	1-10	67	8
Travel	81-90	57	9
Travel	90+	54	10

Ce sont les personnes âgées qui consacrent le plus de temps aux médias, le top 3 étant occupé par des individus de 71 à plus de 90 ans. Ces mêmes individus, en revanche, ne se déplacent que très peu (7e, 9e, 10e positions).

Ce sont les 21-30 ans qui sont les plus mobiles sur une journée. Pourtant, étonnamment, ces derniers sont moins sportifs (8ème place) que les 61-70 ans, qui y consacrent un peu plus de 2h par jour. On observe bien ces tendances sur le graphique.

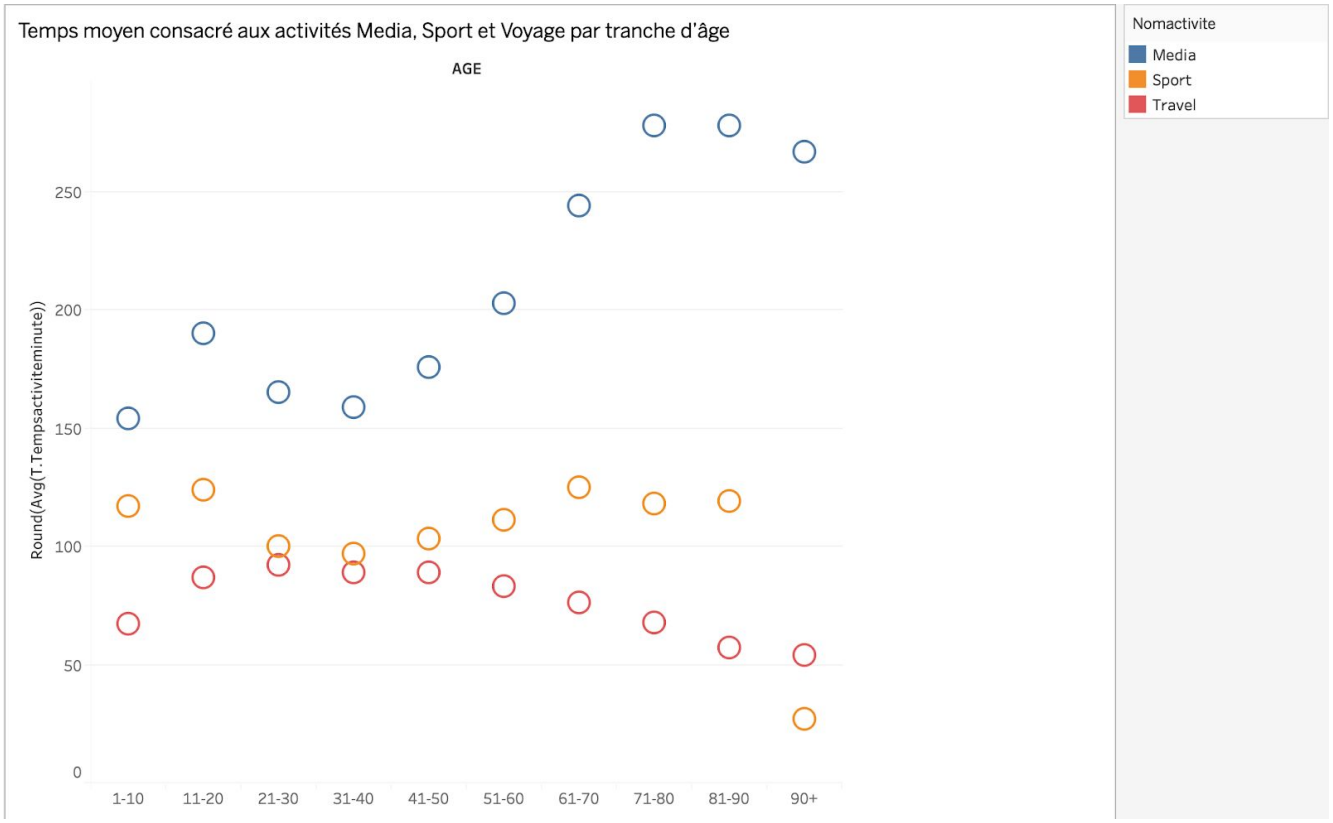


Figure 12. *Graphique représentant les résultats de la requête 7*

Requête 8 (opérateurs GROUP BY GROUPING SETS, GROUPING)

Temps moyen consacré aux activités *Loisirs* et *Médias* pour un individu donné, lorsque celui possède un ordinateur ou un véhicule

```
SELECT a.nomActivite AS Activite, f.aOrdinateur AS A_ordinateur,
f.aVehicule AS A_vehicule, AVG(t.tempsActiviteMinute) AS Temps,
GROUPING(a.nomActivite) AS Grouping_Activite, GROUPING(f.aOrdinateur) AS
Grouping_Ordi, GROUPING(f.aVehicule) AS Grouping_Vehicule
FROM tableFaits t, Personne pe, Activite a, Foyer f
WHERE (a.nomActivite = 'HomeLeisure' OR a.nomActivite = 'Media')
AND a.idActivite = t.idActivite
AND pe.idPersonne = t.idPersonne
AND f.idFoyer = t.idFoyer
GROUP BY ROLLUP a.nomActivite, f.aOrdinateur, f.aVehicule
```

Activite	A_ordinateur	A_vehicule	Temps	Grouping_ Activite	Grouping_ ordi	Grouping_ vehicule
Media			191	0	0	0
Media		18	191	0	0	1
Media	Non		272	0	0	0
Media	Non	Non	217	0	0	0
Media	Non	nan	197	0	0	0
Media	Non	1 voiture/motocycle	172	0	0	0
Media	Non	2+voiture/motocyle	151	0	0	0
Media	Non	Véhicules non-motorisés	190	0	0	0
Media	Non		179	0	0	1
...	...	...	...	...	...	...
Media			196	0	1	1
HomeLeisure			157	0	0	0
HomeLeisure			157	0	0	1
HomeLeisure	...	...	...	...	...	...
HomeLeisure	Oui		107	0	0	0
HomeLeisure	Oui	Non	109	0	0	0
HomeLeisure	Oui	nan	106	0	0	0
HomeLeisure	Oui	1voiture/motocycle	123	0	0	0
HomeLeisure	Oui	2+voiture/motocyle	107	0	0	0
HomeLeisure	Oui	Véhicules non-motorisés	151	0	0	0
HomeLeisure	Oui		115	0	0	1
HomeLeisure			134	0	1	1
HomeLeisure			169	1	1	1

Les personnes ne possédant ni d'ordinateur et ni de véhicule semblent passer plus de temps devant les médias tandis que celles qui ne possèdent pas d'ordinateur mais plus de deux voitures (ou motocyles) en consacrent beaucoup moins. De même, les individus avec ordinateur et plus de deux voitures (ou motocyles) consacreront moins de temps à leurs loisirs comparés aux propriétaires de véhicules non-motorisés.

Requête 9 (opérateurs GROUP BY, RANK)

Top des activités auxquelles un individu consacre le plus de temps, par région habitée (Europe, Amérique du Nord, Moyen-Orient), hors activités *Travail*, *Repos*, *Pause Travail* et *Études*

```
SELECT a.nomActivite AS Activite, p.Region AS Region,
AVG(t.tempsActiviteMinute) AS Temps,
RANK() OVER (ORDER BY AVG(t.tempsActiviteMinute) DESC) AS Rang
FROM tableFaits t, Activite a, Pays p
WHERE  a.idActivite = t.idActivite
AND p.idPays = t.idPays
AND a.nomactivite != 'Work' AND a.nomactivite != 'WorkBreak'
AND a.nomactivite != 'School' AND a.nomactivite != 'Sleep'
GROUP BY a.nomActivite, p.Region;
```

Activite	Region	Temps	Rang
Media	Amérique du Nord	241	1
Media	Proche-Orient	191	2
Tâches Ménagères	Europe	189	3
Media	Europe	182	4
Tâches Ménagères	Proche-Orient	172	5
...	...	...	...
Courses Alimentaires	Proche-Orient	60	29
Courses Alimentaires	Europe	58	30
Soins Personnels	Amérique du Nord	56	31
Cuisine	Amérique du Nord	53	32
Soins Personnels	Proche-Orient	41	33

En Amérique du Nord comme au Moyen-Orient, l'activité la plus chronophage est celle liée aux médias. Les Européens consacrent quant à eux plus de temps aux tâches ménagères. À l'inverse, les activités "Cuisine", "Soins personnels" et "Courses alimentaires" sont celles auxquelles les habitants d'Amérique du Nord, du Moyen-Orient et d'Europe consacrent respectivement le moins de temps.

Requête 10 (opérateurs GROUP BY, NTILE)

Activités auxquelles les personnes handicapées s'adonnent le plus, tous pays confondus, hors activités *Travail*, *Repos*, *Pause Travail* et *Études*

```
SELECT pe.age AS Age, a.nomActivite AS Activite, NTILE(4) OVER (ORDER BY
AVG(t.tempsActiviteSoir) DESC) AS Quartile
FROM tableFaits t, Activite a, Personne pe
WHERE a.idActivite = t.idActivite
AND pe.idPersonne = t.idPersonne AND pe.aHandicap = 'Oui'
AND a.nomActivite != 'Work' AND a.nomActivite != 'WorkBreak'
AND a.nomActivite != 'School' AND a.nomActivite != 'Sleep'
GROUP BY pe.age
```

Age	Activite	Quartile
26	Religion	1
19	Soins apportés aux enfants	1
72	Média	1
79	Média	1
62	Média	1
...	...	...
37	Religion	2
12	Loisirs	2
48	Tâches ménagères	2
75	Tâches ménagères	2
52	Tâches ménagères	2
...	...	...

Age	Activite	Quartile
11	Travel	3
70	Religion	3
88	Travel	3
15	Sport	3
39	Travel	3
...	...	...
88	Soins apportés aux enfants	4
84	Cooking	4
52	Grocery	4
83	Travel	4
51	Grocery	4
...	...	...

5. Conclusion

Ce projet s'est avéré extrêmement enrichissant, de par sa complétude d'abord, puisqu'il nous a permis de mettre en oeuvre toutes les étapes liées à la conception d'un entrepôt de données, incluant la construction de notre propre dataset. Les étapes liées au prétraitement et à l'intégration des données sont celles auxquelles nous avons ici consacré le plus de temps, de par les nombreux problèmes que nous y avons rencontrés. Une mauvaise coordination et un manque de recul sur ces étapes nous ont parfois ralentis, voire conduits à devoir renoncer à certains attributs, qui auraient pourtant permis d'enrichir davantage notre analyse. Ces quelques erreurs de parcours sont toutefois à nuancer, de par le caractère ambitieux de notre projet (données abondantes, codifiées, parfois non homogénéisées). L'idée d'avoir pu travailler sur une sujet au coeur de notre vie quotidienne s'est avérée très séduisante. Les corrélations d'ordre sociales/sociologiques que nous avons pu dégager au travers de nos requêtes nous ont permis de mettre en évidence des habitudes de vie, c'est-à-dire retracer, en quelque sorte, le rapport de l'homme au temps, à la fois sur une échelle spatiale et temporelle.

ANNEXE

Annexe 1 : *Table des faits*

Table des Faits	
Attribut	**Description**
idTableFaits	Clé primaire
idPersonne	Clé étrangère de la table Personne
idFoyer	Clé étrangère de la table Foyer
idActivite	Clé étrangère de la table Activite
idPays	Clé étrangère de la table Pays
idDate	Clé étrangère de la table Date
idSituationPro	Clé étrangère de la table Situation professionnelle
tempsActiviteMinute	Temps passé pour une activité en minutes sur la journée
tempsActiviteMatin	Temps passé pour une activité durant la plage horaire MATIN
tempsActiviteApresMidi	Temps passé pour une activité durant la plage horaire APRÈS-MIDI
tempsActiviteSoir	Temps passé pour une activité durant la plage horaire SOIR
tempsActiviteNuit	Temps passé pour une activité durant la plage horaire NUIT

Annexe 2 : *Tables des dimensions*

Personne	
Attribut	**Description**
idPersonne	Clé primaire de la table Personne
age	Âge de la personne
sexe	Sexe de la personne
sante	État de santé de la personne
situationMaritale	Situation maritale de la personne
estCitoyen	Indique si la personne est citoyen(ne) du pays dans lequel elle a été interrogée
niveauEtudes	Indique le niveau d'éducation de la personne
aHandicap	Indique si la personne souffre d'un handicap

Pays	
Attribut	**Description**
idPays	Clé primaire de la table Pays
nomPays	Nom du pays
region	Région du pays

Activité	
Attribut	**Description**
idActivite	Clé primaire de la table Activité
nomActivite	Nom de l'activité

Date	
Variable	**Description**
idDate	Clé primaire de la table Date
annee	Année

Foyer	
Attribut	**Description**
idFoyer	Clé primaire de la table Foyer
taille	Nombre de personnes dans le foyer
lieuResidence	Type de lieu de résidence (*urbain*, *rural*)
aOrdinateur	Indique si le foyer a un ordinateur ou accès à internet
aVehicule	Indique si le foyer a un véhicule
ageBenjamin	Indique l'âge de l'enfant le plus jeune du foyer

Situation Professionnelle	
Attribut	**Description**
idSituationProfessionnelle	Clé primaire de la table Situation Professionnelle
situation	Indique la situation professionnelle
revenu	Revenu de la personne (par tranche)
secteur	Secteur (*public ou privé*)
typeEmploi	Indique si la personne a un emploi ou non
typeEmploiConjoint	Indique si le/a conjoint(e) de la personne a un emploi ou non

Annexe 3 : *Ensemble des valeurs prises par les attribut catégoriques (non traduits)*

COUNTRY
NL Netherlands
ES Spain
UK United Kingdom
US USA
FR France
AT Austria
FI Finland
CA Canada
IL Israel
HU Hungary
IT Italy

VEHICLE
00 No
01 Animal only
02 Non-motorized vehicle
03 1 car/motocycles
04 2+ cars/motocycles
-8 Missing
-9 Not applicable/not asked

COMPUTER
00 No
01 Yes
-7 Not applicable/not asked

URBAN
00 Urban/Suburban
01 Rural/Semi-rural
-7 Not applicable/not asked
-8 Missing

SEX
01 Man
02 Woman
-8 Missing

CITIZEN
01 No
02 Yes
-7 Not applicable/not asked
-8 Missing

EDTRY
00 Uncompleted secondary or less
01 Completed secondary
02 Above secondary education
-7 Not applicable/not asked
-8 Missing
-9 Could not be created

CIVSTAT
01 In couple (married/cohabit/civil partnership)
02 Not in couple
-8 Missing

EMPSTAT
01 Full-time
02 Part-time
03 Unknown job hours
04 Not in paid work
-7 Not asked of this diarist
-8 Missing

EMPSP
01 Full-time
02 Part-time
03 Unknown job hours
04 Not in paid work
-7 Not asked of this diarist
-8 Missing

SECTOR
01 Public sector
02 Private sector
-7 Not applicable/not asked
-8 Missing

HEALTH
00 Poor
01 Fair
02 Good
03 Very Good
-7 Not applicable/not asked
-8 Missing
-9 Could not be created

DISAB
00 No
01 Yes
-7 Not applicable/not asked
-8 Missing
-9 Could not be created

Annexe 4 : *Correspondance des attribut catégoriques du dataset et des attributs catégoriques des tables des faits et des dimensions*

Attributs du dataset	**Attributs tables de faits & dimensions**	
	Nom attribut	**Nom table**
COUNTRY	*nomPays*	Pays
VEHICLE	*aVehicule*	Foyer
COMPUTER	*aOrdinateur*	Foyer
URBAN	*lieuResidence*	Foyer
SEX	*sexe*	Personne
CITIZEN	*estCitoyen*	Personne
CIVSTAT	*situationMaritale*	Personne
EDTRY	*niveauEtudes*	Personne
HEALTH	*sante*	Personne
DISAB	*aHandicap*	Personne
EMPSTAT	*typeEmploi*	SituationPro
EMPSP	*typeEmploiConjoint*	SituationPro
SECTOR	*secteur*	SituationPro
INCOME	*revenu*	SituationPro